One Frog, Two Frog, Small Frog, Big Frog: A Glimpse at Amphibian Body Size

Kevin Nguyen

Abstract

AmphiBIO is a large dataset collected by very talented individuals, and was created for the purpose of studying this diverse animal group. This analysis was motivated in hopes of being able to better understand what factors influence an amphibians body size. The extensive data set was narrowed down to two categorical and two quantitative variable which were all randomly sampled. Then two one-way ANOVA tests were performed using the categorical variables, habitat type, and diel activity and a linear regression to determine body size from an amphibians reproductive output was performed to determine if any predictability could be observed. There were no significant results found from the two tests, habitat type and diel activity do not have any significant effect on an amphibian's body size. Furthermore, reproductive output can not be used to predict an amphibian's body size. These results only is able to display that analysis of only abiotic factors and behavioral patterns are not sufficient to provide a thorough analysis on amphibians.

Introduction

This data set compiled primarily by researchers at Federal University of Rio Grande do Norte National, Autonomous University of Mexico, and University at Montgomery. The data set AmphiBIO is a compilation of ecological traits of amphibians across the world. The data was gathered by analyzing over 1,500 scientific literature sources for over 6,500 species. It is comprised of a total of 6776 entries and 17 traits in hopes of allowing for a large scale analyses in ecology,evolution, and conservation of amphibians.

Though there are a variety of traits that can be utilized to study, from this data set, I want to see if either the habitat or diel activity have any significant effect on the body size of amphibians. Furthermore, I would like to see if we can predict the amphibian's body size based on its reproductive output. I predict that the habitat and diel activity will have a significant effect on amphibian body size, however I hypothesize that reproductive output of the amphibian will not be able to predict its body size. A note to make is that due to the large size of the dataset, a random sample of 200 with a saved seed was created to perform this analysis so that it could be more manageable.

Exploratory Data Analysis

Histogram

To see the distribution of the numerical data, histograms were created. Both body size and reproductive output had a large right skew suggesting non-normality, and is further confirmed by a Shapiro-Wilk test with p values of 3.087e-15 and 2.2e-16 respectively which is less than 0.05 and rejects the null hypothesis of normality. A log transformation was done and body size had a much more normal distribution and p value of 0.1945 allowing the

assumptions of normality to be met. A log transformation for the reproductive output did not produce any noticeable changes and can be ignored since that is not the response variable that is being analyzed.

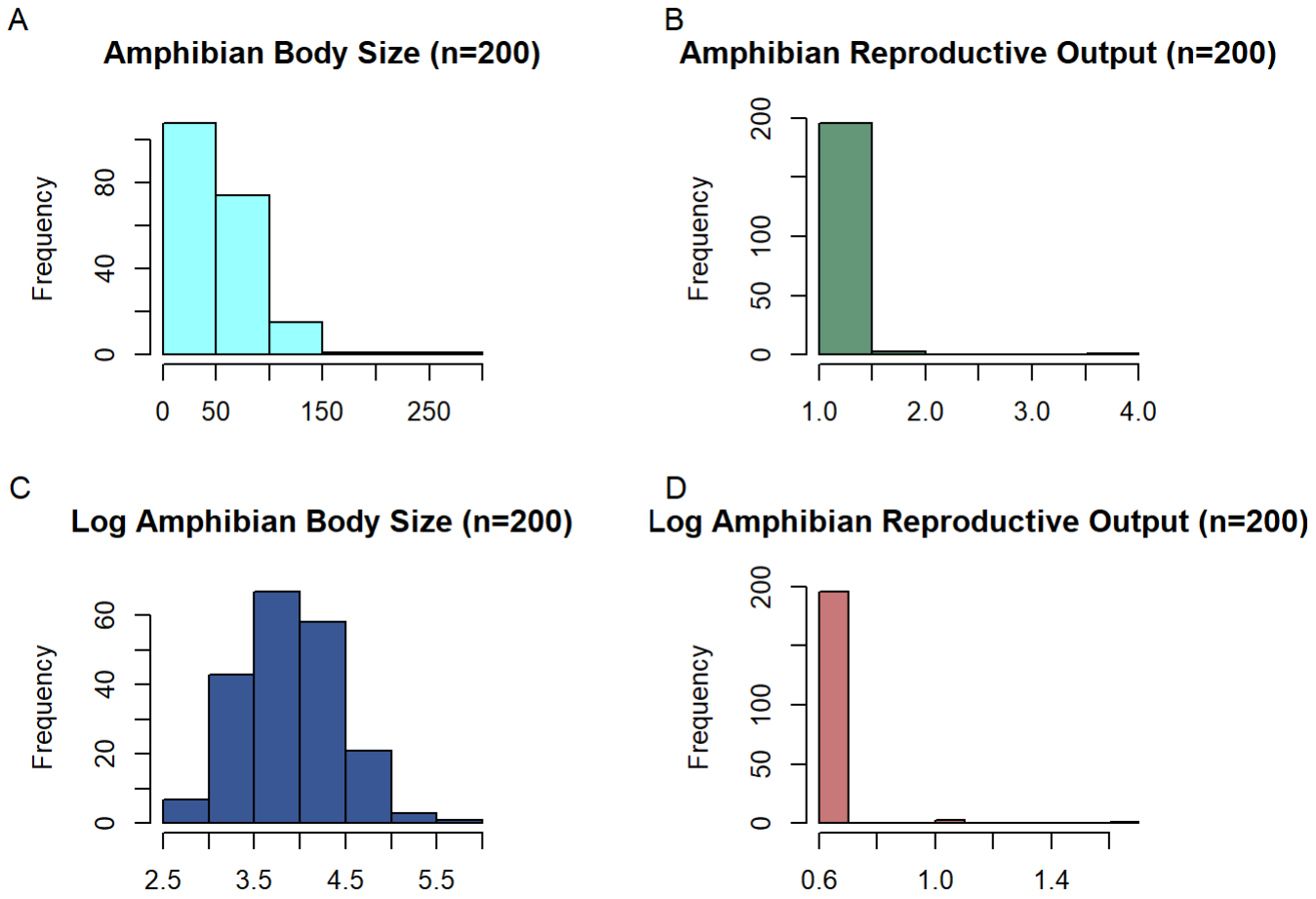


Figure 1: A) Histogram of amphibian body size that has a large right skew and does not follow normal distribution B) Histogram of amphibian reproductive output that has a large right skew, almost binary in distribution and does not follow normal distribution. C) Histogram of log transformed data of amphibian body size that appears to have a bell shape curve and can be considered approximately normal with the Central Limit Theorem as well. D) Histogram of log transformed data of amphibian reproductive output has a large right skew and does not follow normal distribution

Boxplot

Boxplots were utilized to see how the amphibian body size differs across both habitat and diel. When comparing body size across habitats, there is a larger range in size in amphibians found in terrestrial habitats. The medians across all habitats are relatively the same and consistent with one another. The Fossorial and Terrestrial habitats both have two outliers but can still be argued to be normal due to the Central Limit Theorem. When comparing across different diel activity, there is a greater variability in body size for amphibians with a nocturnal diel activity. The medians across all diel activity are relatively the same and consistent with one another. Both amphibians with diurnal and nocturnal have 1 and 2 outliers respectively, but can still be argued to fulfill normality due to the Central Limit Theorem.

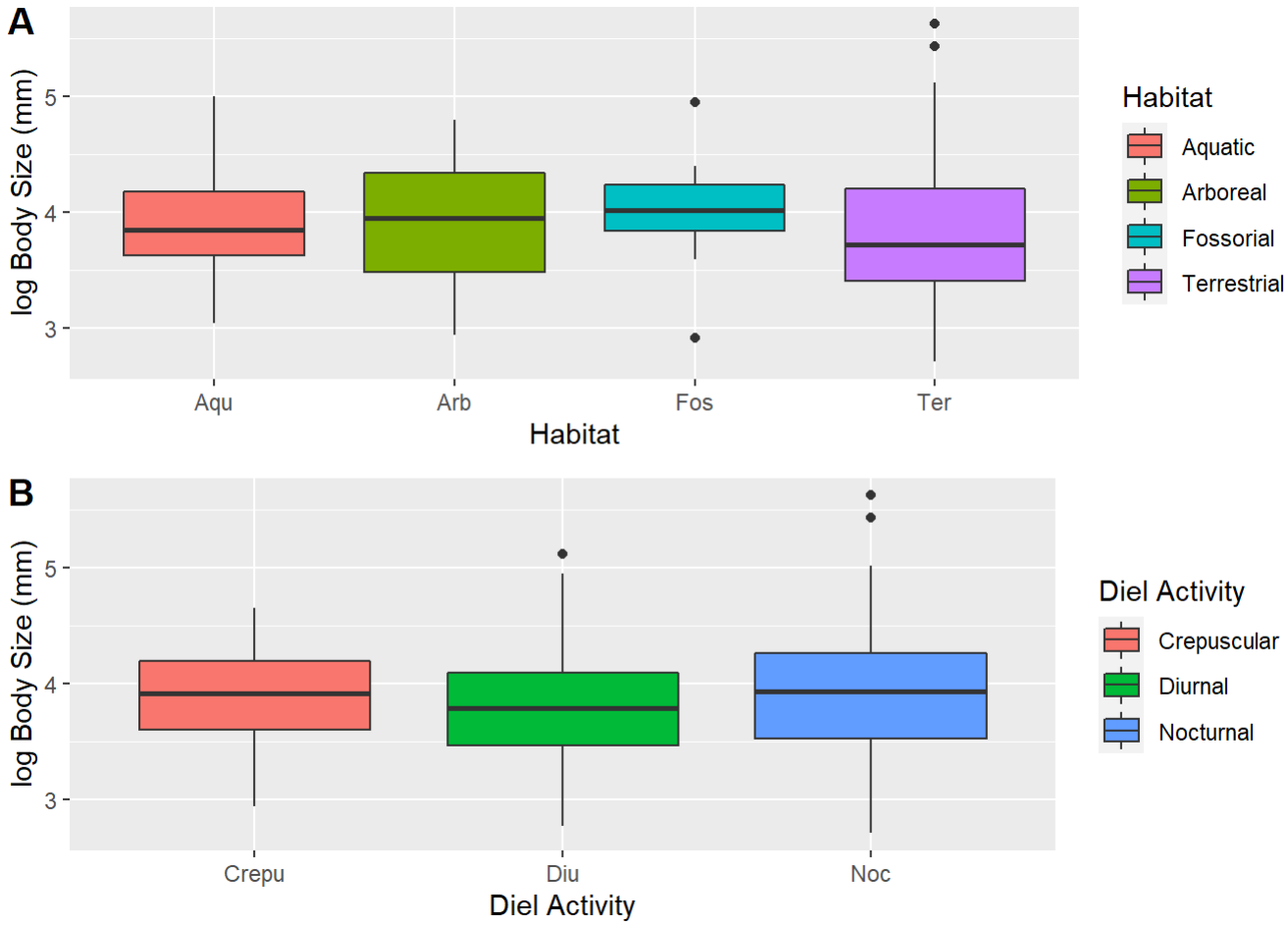


Figure 2: A) Boxplot of log transformed data of body size across four different habitats. B) Box plot of log transformed data of body size across 3 different diel activities. Amphibians with nocturnal diel activity have the most outliers but is still approximately normal according to the Central Limit Theorem

Statistical Methods

ANOVA

One of the statistical methods being used is an ANOVA test which is simply an analysis of variance. It can be used to determine if the populations of various groups are statistically different from one another. For this data set a two one way ANOVA tests, the dependent variable is the body size, while the two independent variables are the habitat type and type of diel activity. The habitat has 4 levels being aquatic, arboreal, fossorial, and terrestrial. Diel Activity has 3 levels of being crepuscular, diurnal and nocturnal.

Before the ANOVA test can be performed, it must first fulfill three assumptions.

One of the assumptions is that the samples are randomly sampled which is met by the random sample generation from the entire data set.

The next assumption is that the residuals must be normally distributed which can be tested by looking at the “Normal Q-Q plot” and “Residual vs fitted” plot. A residual is the difference between the experimental and theoretical data which is signified by the red line in the residual plot. From our test we can see that in the “Normal QQ” plot the majority of the data points follow the trendline and in the residual plot, there is no distinct pattern formed about the 0 line, so this assumption is also met.

The final assumption is that there must be an equality of variance which is observed in the boxplots, and therefore this assumption is also met. The variance is determined by the difference between the collected datapoints and the mean of the dataset used.

The null hypothesis of an ANOVA is that a p-value greater than 0.05 is found, then the independent variables (habitat & diel activity) do not have a significant effect on the amphibian's body size where the mean across all levels are equal to one another. The alternative hypothesis is that if the p-value is less than 0.05, then at least one of the independent variables has a significant effect on the amphibian's body size where at least one of the means across all levels are not equal to one another.

Linear Regression

To perform a predictive analysis, a linear regression test is carried out. In this study, performing a linear regression will help us understand if it is possible to predict an amphibians body size based on its reproductive output.

Before a linear regression model can be made, it must first fulfill three assumptions as well.

One assumption is that the data come from a random sample, which is met by the random sampling done on the data set.

Another assumption is that there must be an equality of variance which can be tested by plotting the residuals and observing the "Residuals vs. Fitted" plot. There appears to be no distinct patterns about the 0 line so this assumption is also met.

Finally there must be normality of the residuals. This is tested and met by the "Normal Q-Q" plot, since when looking at the plot, it the data points closely follow the trendline. Therefore the residuals can be assumed to be normal.

The null hypothesis for the linear regression is that if the p-value is greater than 0.05, then an amphibian's body size cannot be predicted by its reproductive output. The alternative hypothesis is that if the p-value is less than 0.05, then the amphibian's body size can be predicted by its reproductive output.

Results

Anova

Each ANOVA test resulted in one p value which is compared to the alpha level of 0.05. For the ANOVA test on the different habitats, we find a p-value of 0.738. Since this value is greater than 0.05, I am unable to reject the null hypothesis and find that there is no significant differences between the body sizes of amphibians across different habitats.

The ANOVA assessment on the diel activity procured a p value of 0.547, which is also greater than 0.05, and therefore am unable to reject the null hypothesis and conclude that there is no significant differences between the body sizes of amphibians across different diel activity types.

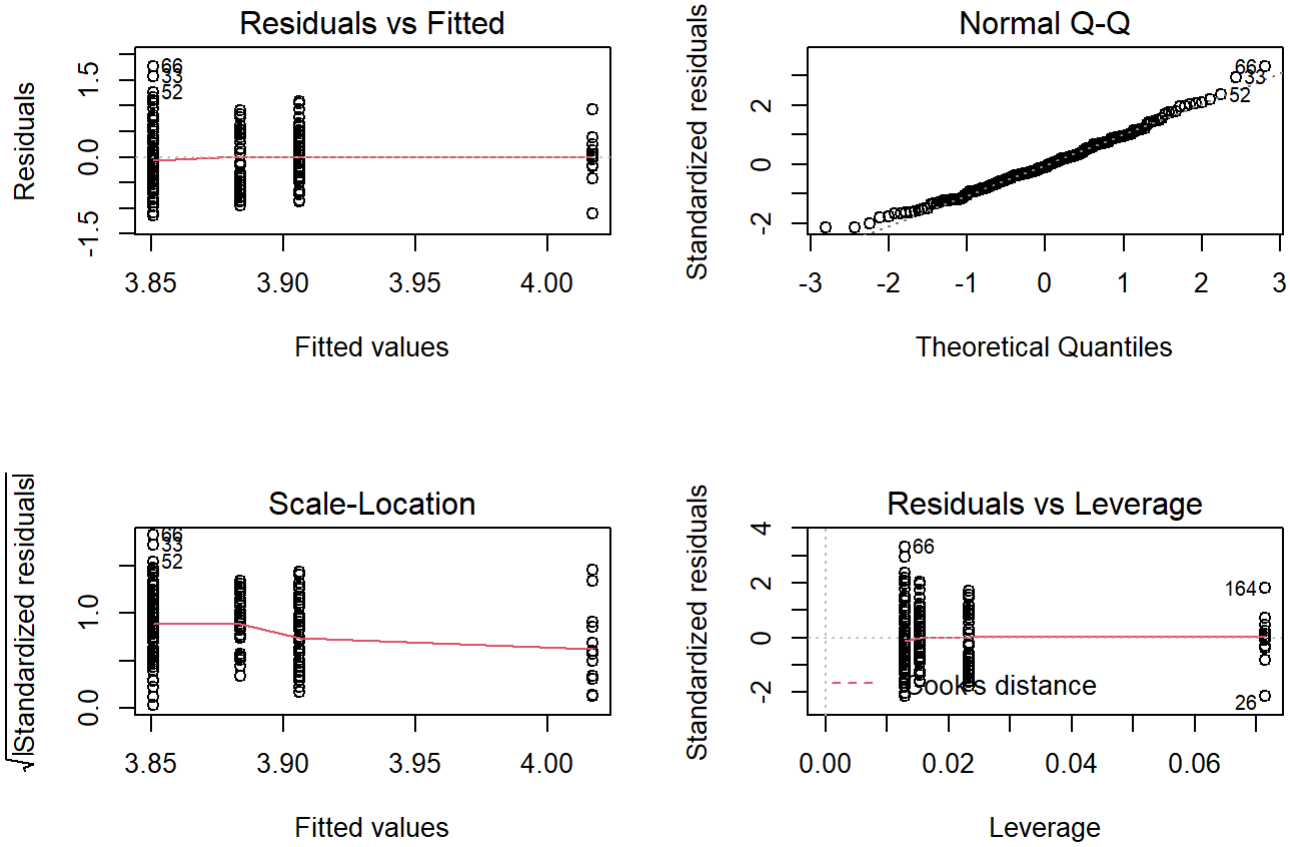


Figure 3: Residual plots for the ANOVA test for the Habitats. The residuals vs. fitted plot appears homoskedasic which fulfills the assumption of equality of variances. The normal Q-Q plot exhibits the normality due the the majority of the data points following the trendline fufilling the assumption of normality.

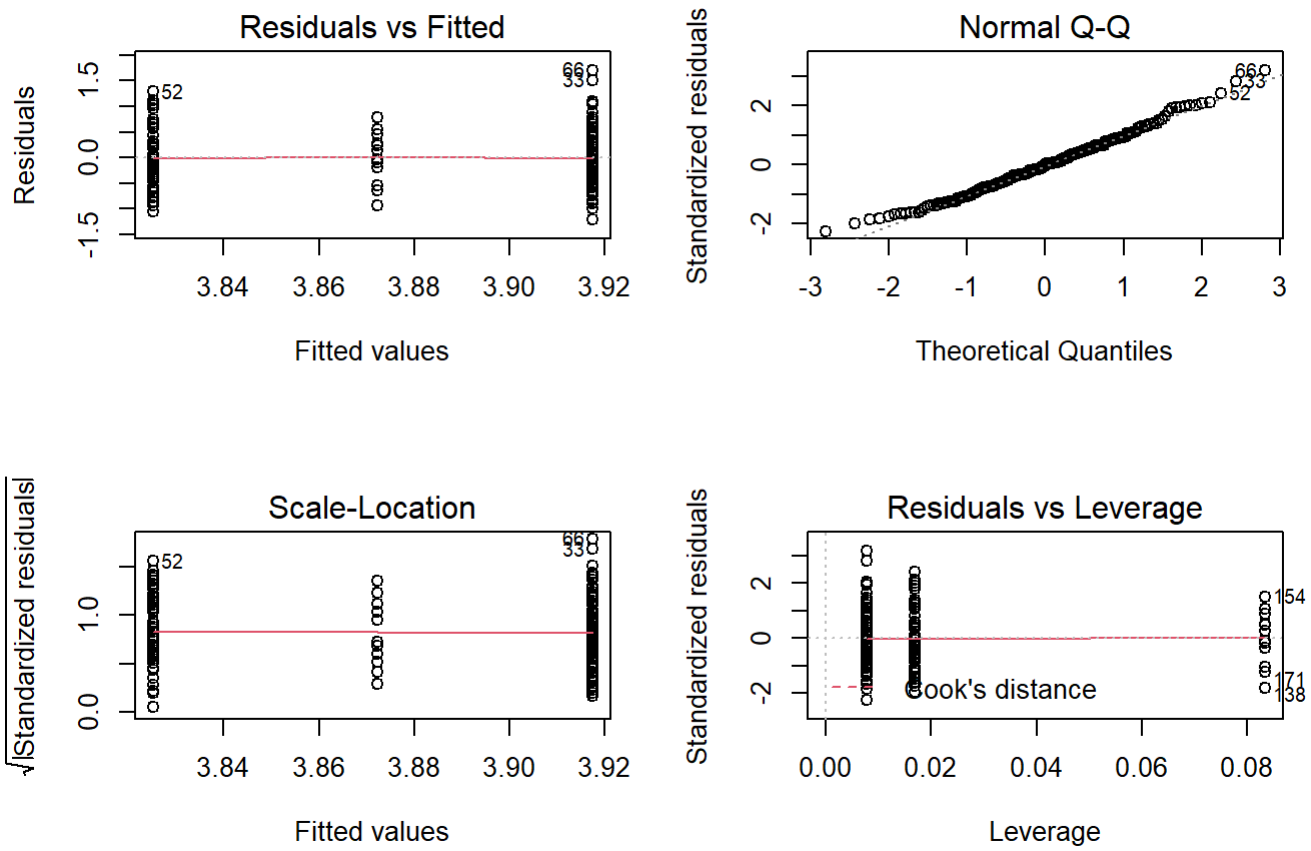


Figure 4:Residual plots for the ANOVA test for Diel Activity. The residuals vs. fitted plot appears homoskedasic which fulfills the assumption of equality of variances. The normal Q-Q plot exhibits the normality due the the majority of the data points following the trendline fulfilling the assumption of normality.

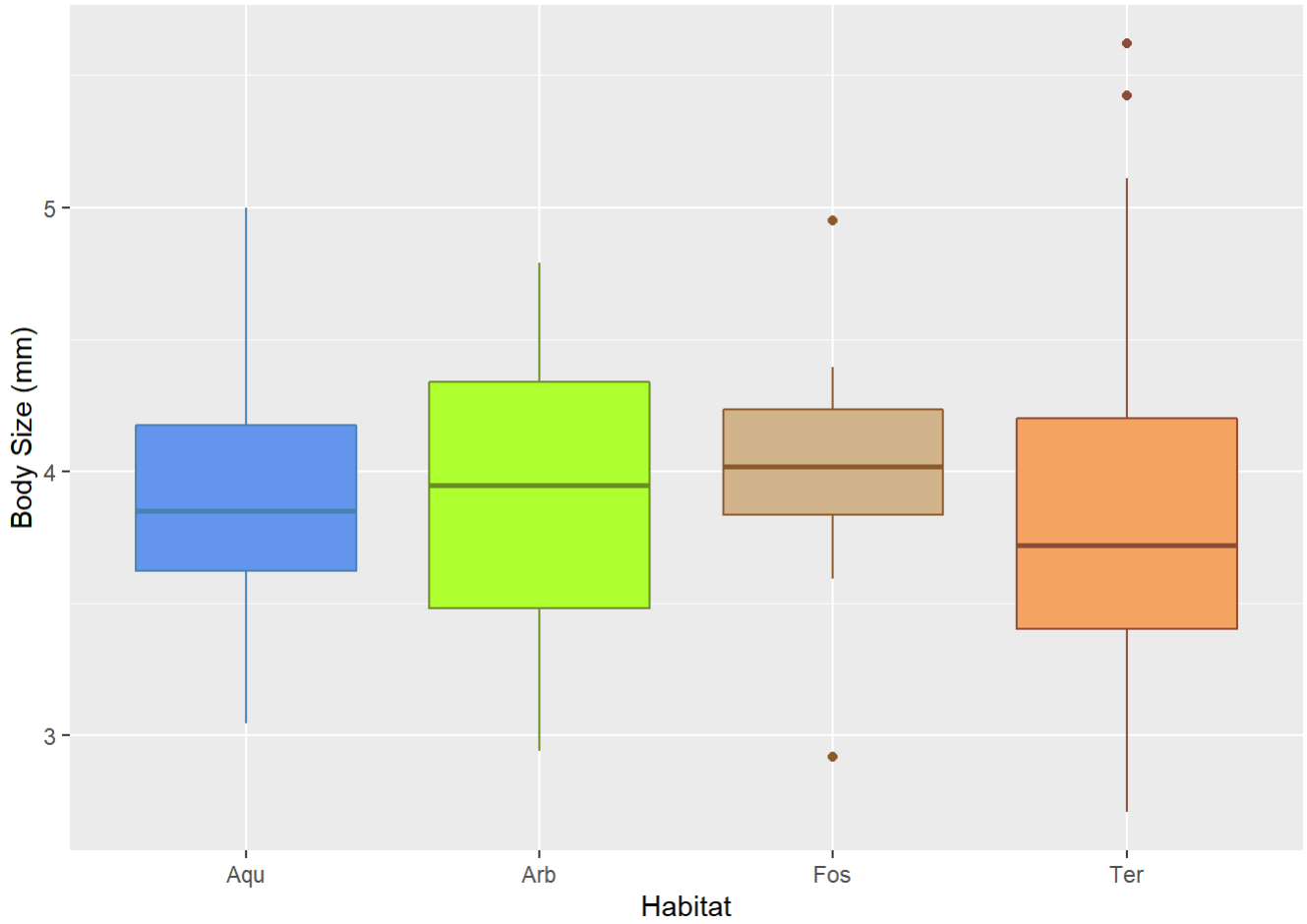


Figure 5:A visualization of the results of the ANOVA test for Habitats. The plot shows that there is no significant difference between the body sizes of amphibians across the different habitats.

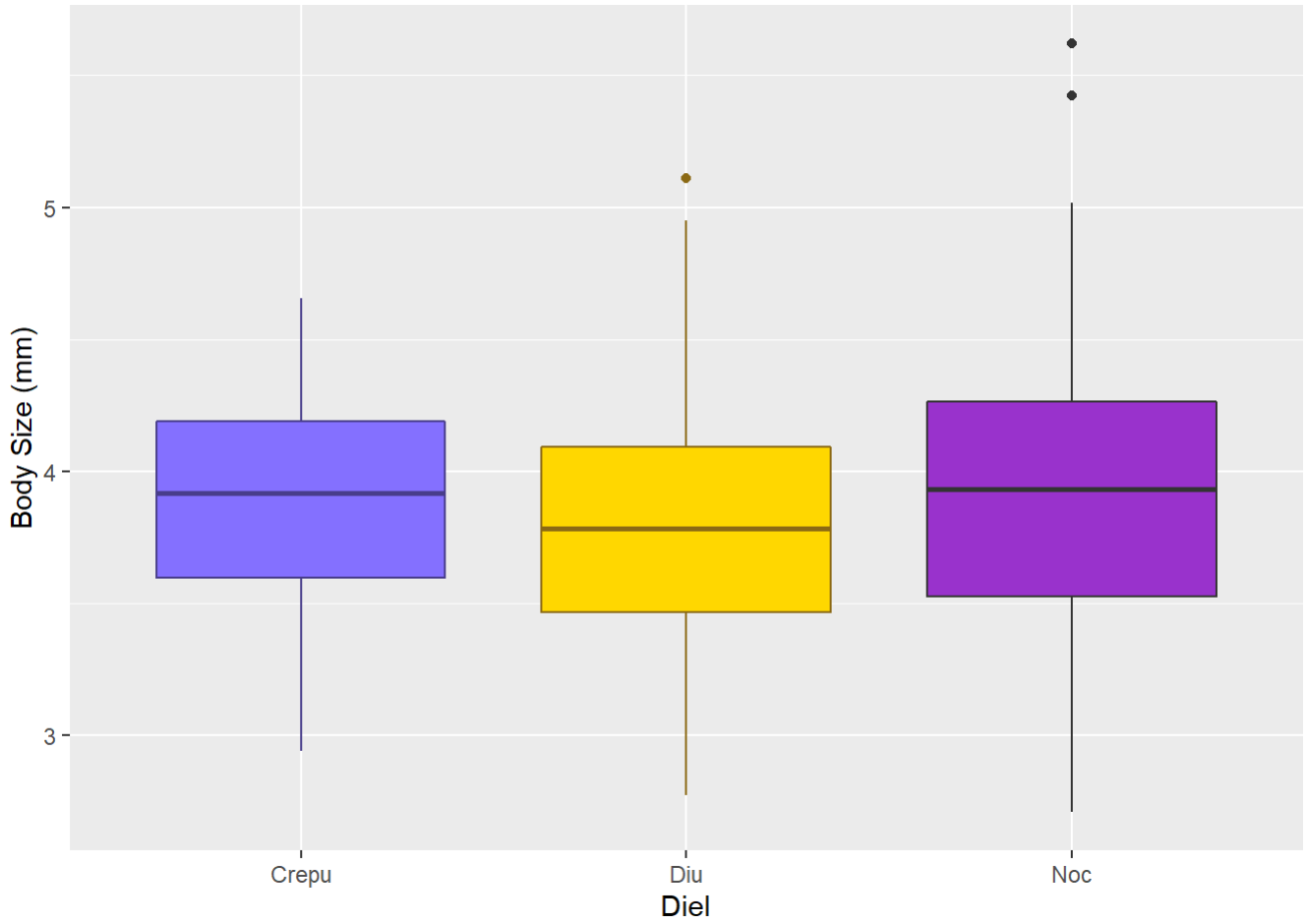


Figure 6:A visualization of the results of the ANOVA test for Diel Activity. The plot shows that there is no significant difference between the body sizes of amphibians across the different habitats.

Linear Regression

From the Linear Regression model, we find a p-value of 0.12. Since it is greater than 0.05, we are unable to reject the null hypothesis and find that reproductive output is not a good indicator to predict an amphibian's body size.

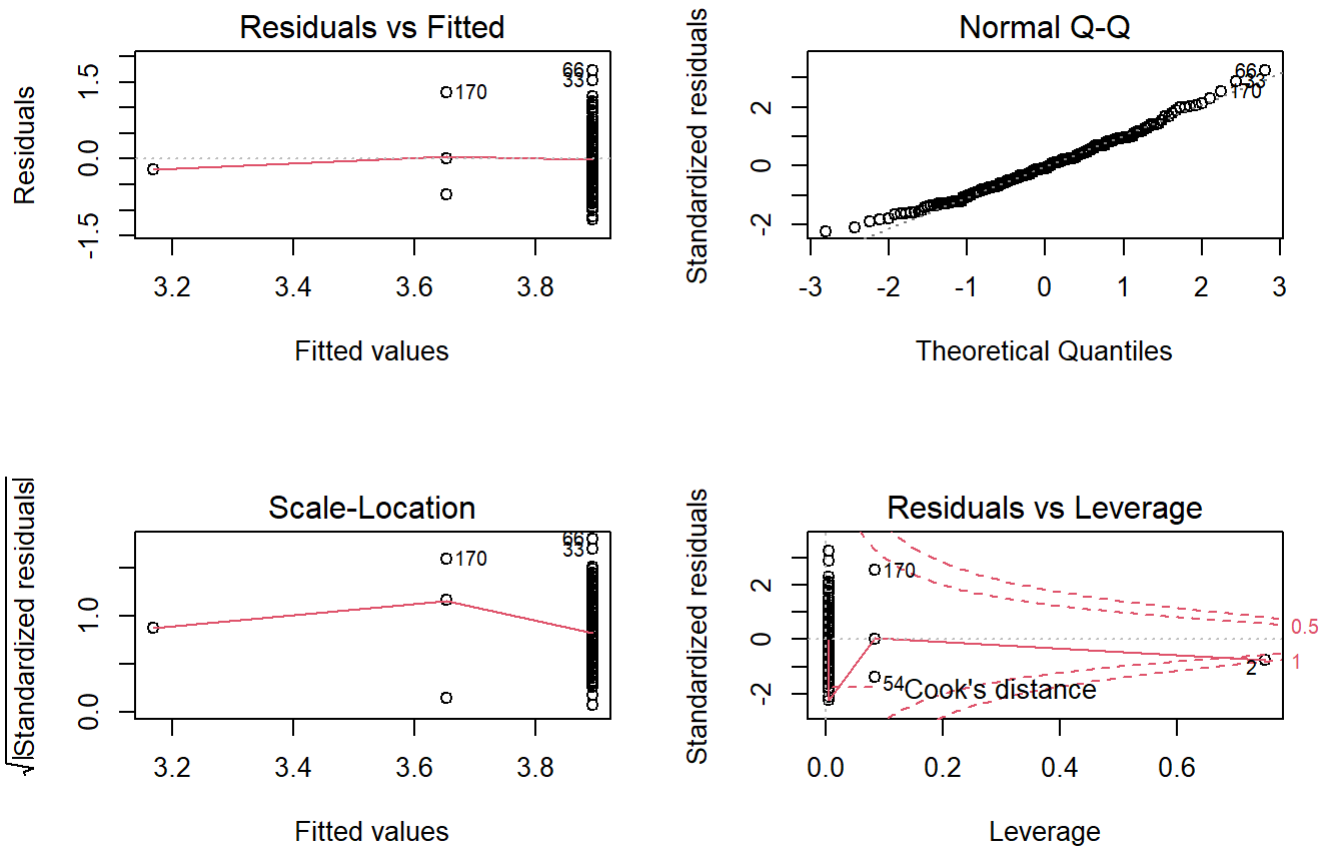


Figure 7:Residual plots for the linear regression. The residuals vs. fitted plot appears homoskedasic which fulfills the assumption of equality of variances. The normal Q-Q plot exhibits the normality due the the majority of the data points following the trendline fulfilling the assumption of normality.

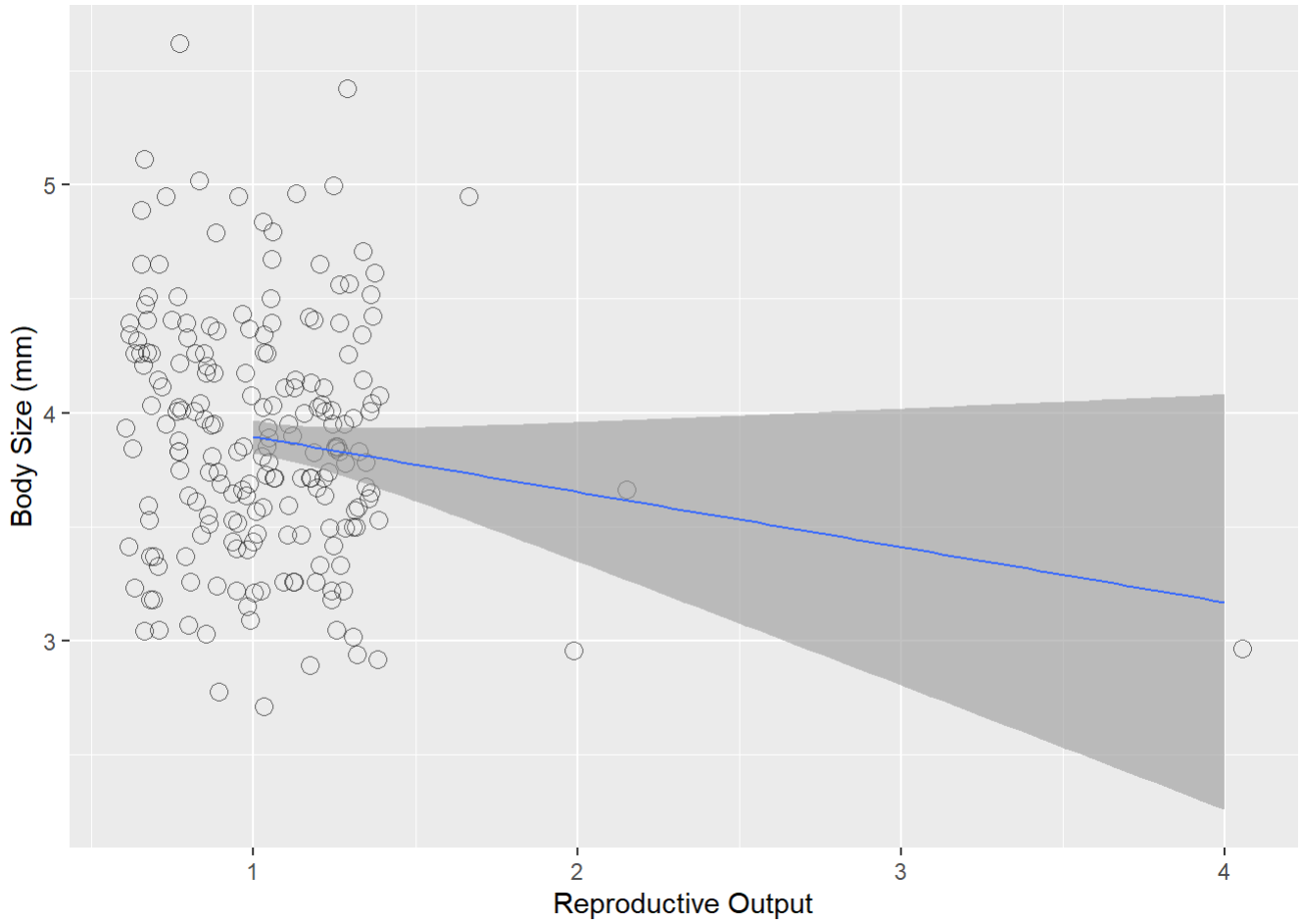


Figure 8: A visualization of the results of the linear regression. The datapoints and trendline for reproductive output are not very closely aligned and appears to be grouped up on the left providing further evidence that reproductive output is not a good predictor for body size

Discussion

The results from our ANOVA tests demonstrated that both habitat type and diel activity type do not appear to significantly contribute to the body size of an amphibian. Though no significant results were found, this allows us to understand that the the factors that influence the sizes of amphibians are much more intertwined. I hypothesized that there would be significant results because I assumed that the different habitats would have different nutrients available to each amphibian, furthermore I also assumed that by having differing diel activity, amphibians would be available to different types of nutrients that could possibly affect their body size. It seems that the majority body sizes of amphibians are conserved across the world despite the habitat type and diel activity type. The limitation in this analysis is that only two variables were observed to compare to an amphibians body size. Perhaps in a future test, more factors should be involved to see if there is something influencing the sizes of the amphibians. Furthermore, the data set is very broad across multiple species, which provides a very broad variation in data so given more time and data, I would focus my assessments on a more select few species instead to understand what could possibly influence body size.

In our linear regression model, our results indicated that reproductive output can not be used to predict an amphibians body size. My hypothesis on this assumption was correct, and makes sense since a reproduction event is not indicative of how the individual is. This test was quite limited in its scope, so perhaps next time I will select features in amphibians that are more similar with one another to provide a much more helpful predictive analysis

The final takeaway in this study is that the amphibians like many other animals are very complex creatures in the ecosystem, and its physical attributes can not simply be explained by a few variables. It is important to understand that there are alot of variables that influence the development of amphibians and there is still much more data to work with and analyze.

References

Oliveira, Brunno Freire; São-Pedro, Vinícius Avelar; Santos-Barrera, Georgina; Penone, Caterina; C. Costa, Gabriel (2017): AmphiBIO_v1. figshare. Dataset. https://doi.org/10.6084/m9.figshare.4644424.v5 (https://doi.org/10.6084/m9.figshare.4644424.v5)

```
citation("psych")
```

```
##
## To cite the psych package in publications use:
##
##   Revelle, W. (2020) psych: Procedures for Personality and
##   Psychological Research, Northwestern University, Evanston, Illinois,
##   USA, https://CRAN.R-project.org/package=psych Version = 2.0.12,.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {psych: Procedures for Psychological, Psychometric, and Personality Research},
##     author = {William Revelle},
##     organization = { Northwestern University},
##     address = { Evanston, Illinois},
##     year = {2020},
##     note = {R package version 2.0.12},
##     url = {https://CRAN.R-project.org/package=psych},
##   }
```

```
citation("car")
```

```
##
## To cite the car package in publications use:
##
##   John Fox and Sanford Weisberg (2019). An {R} Companion to Applied
##   Regression, Third Edition. Thousand Oaks CA: Sage. URL:
##   https://socialsciences.mcmaster.ca/jfox/Books/Companion/
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     title = {An {R} Companion to Applied Regression},
##     edition = {Third},
##     author = {John Fox and Sanford Weisberg},
##     year = {2019},
##     publisher = {Sage},
##     address = {Thousand Oaks {CA}},
##     url = {https://socialsciences.mcmaster.ca/jfox/Books/Companion/},
##   }
```

```
citation("ggplot2")
```

```
##
## To cite ggplot2 in publications, please use:
##
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2016.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2016},
##     isbn = {978-3-319-24277-4},
##     url = {https://ggplot2.tidyverse.org},
##   }
```

```
citation("tidyverse")
```

```
##
##   Wickham et al., (2019). Welcome to the tidyverse. Journal of Open
##   Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Welcome to the {tidyverse}},
##     author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy
D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and
Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kir
ill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske
Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani},
##     year = {2019},
##     journal = {Journal of Open Source Software},
##     volume = {4},
##     number = {43},
##     pages = {1686},
##     doi = {10.21105/joss.01686},
##   }
```

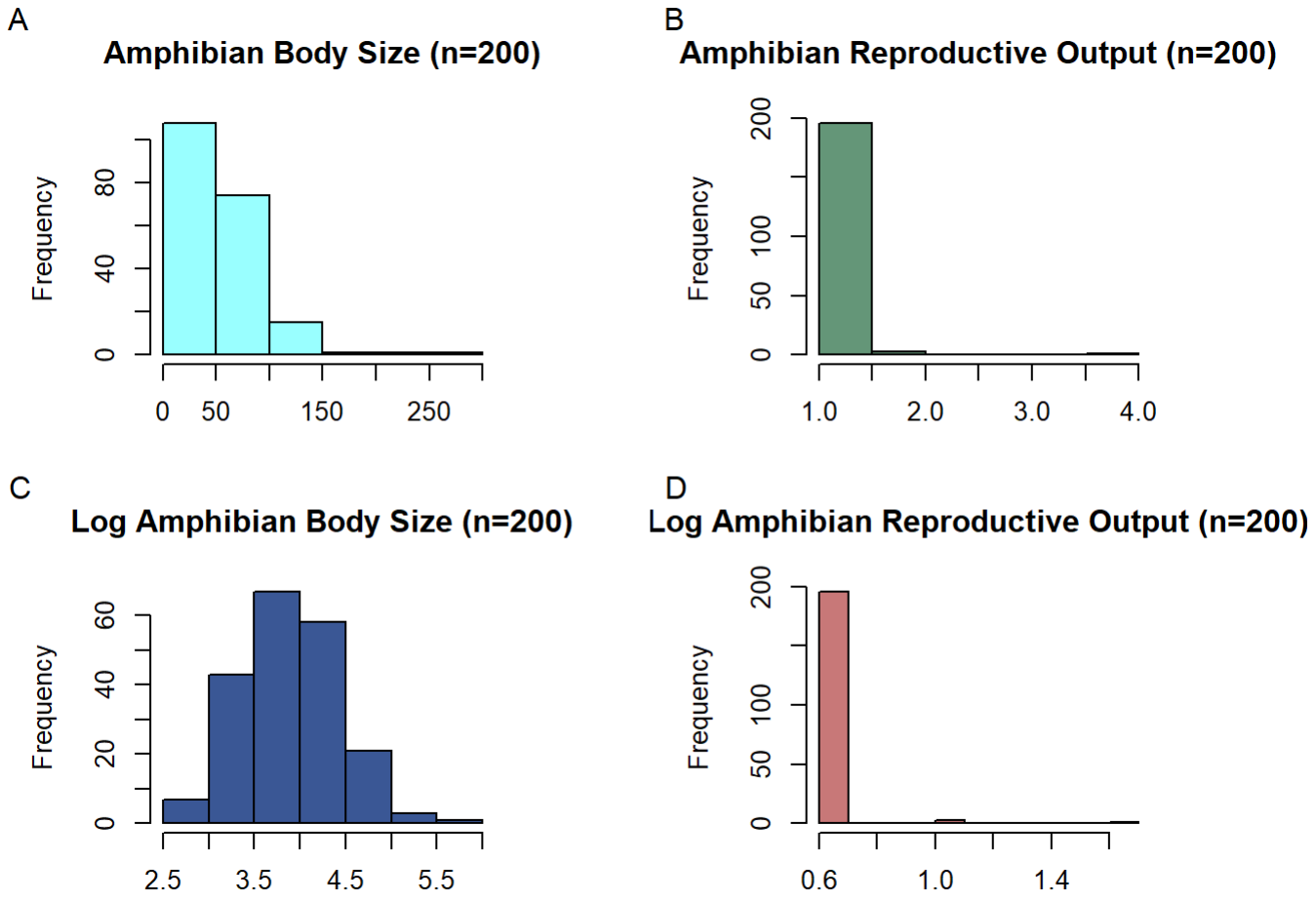
```
citation("ggpubr")
```

```
##
## To cite package 'ggpubr' in publications use:
##
##   Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication
##   Ready Plots. R package version 0.4.0.
##   https://CRAN.R-project.org/package=ggpubr
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {ggpubr: 'ggplot2' Based Publication Ready Plots},
##     author = {Alboukadel Kassambara},
##     year = {2020},
##     note = {R package version 0.4.0},
##     url = {https://CRAN.R-project.org/package=ggpubr},
##   }
```

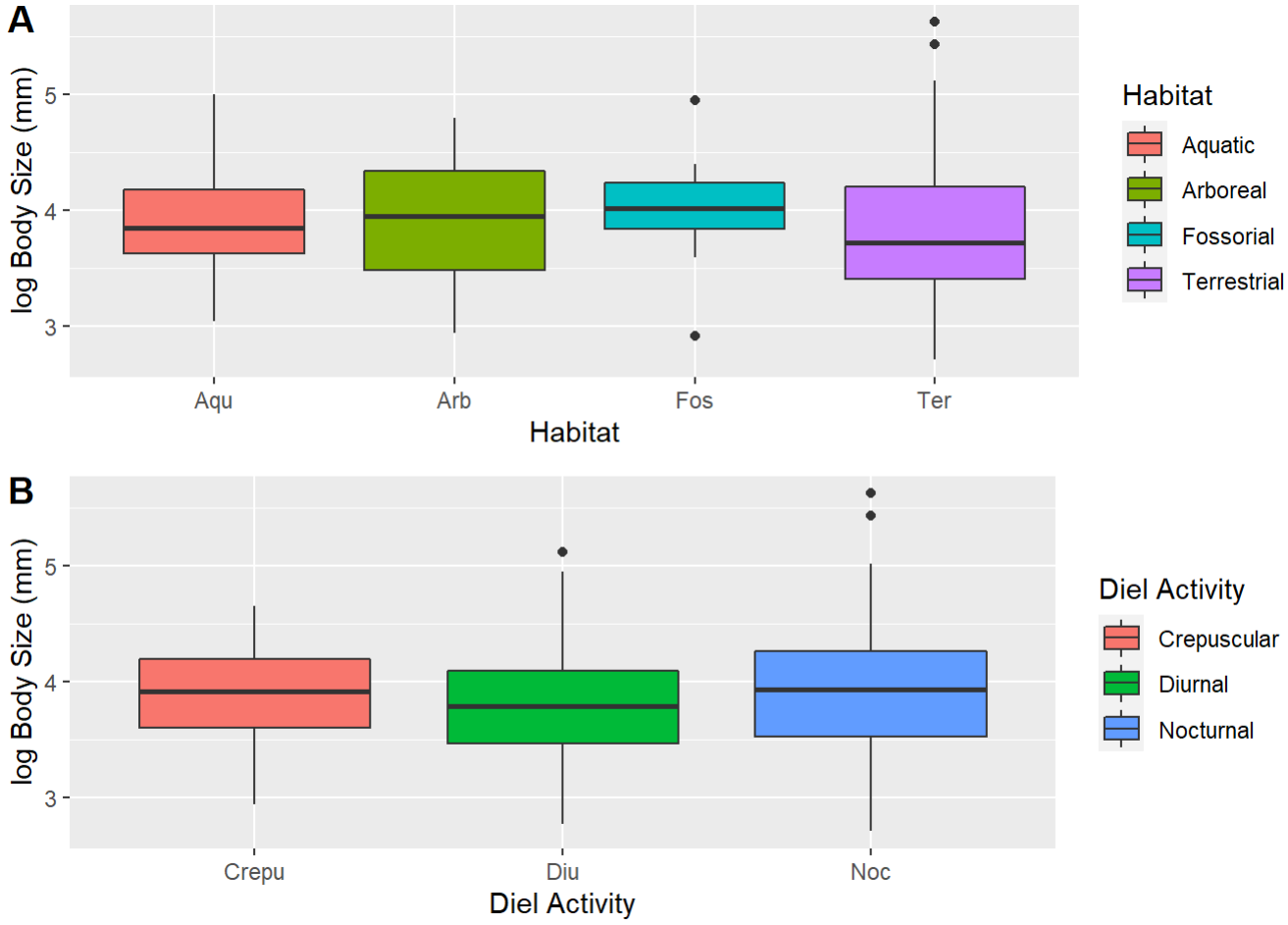
Appendix

```
#Load AmphiBIO data
amphi <- read_csv("AmphiBIO_v1.csv")

#Subsetting Data with Variables We are Focusing On
amphia <- amphi[, colnames(amphi)[c(5:9, 16:18, 26, 32)]]
#Removing NA values
amphib<- amphia[rowSums(is.na(amphia[c("Diu", "Noc", "Crepu")])) != 3, ]
amphic<- amphib[rowSums(is.na(amphib[c("Body_size_mm")])) != 1, ]
amphid<- amphic[rowSums(is.na(amphic[c("Reproductive_output_y")])) != 1, ]

#Pivoting Habitat and Diel
AmphiA <- amphid %>%
  pivot_longer(c(`Fos`, `Ter`, `Aqu`, `Arb`), names_to = "Habitat", values_to = "cases")

AmphiB <- AmphiA %>%
  pivot_longer(c(`Diu`, `Noc`, `Crepu`), names_to = "Diel", values_to = "case")

#Removing Rows with NAs
AmphiC<- AmphiB[rowSums(is.na(AmphiB[c("cases")])) != 1, ]
AmphiD<- AmphiC[rowSums(is.na(AmphiC[c("case")])) != 1, ]

#Final Data Set (removing unneeded columns)
AmphiE <- subset( AmphiD, select = -c(cases,case ) )

#Random Seed
set.seed(120)
  AmphiF <- sample_n(AmphiE,200)

#Converting to Factors
AmphiF$Habitat <- as.factor(AmphiF$Habitat)
AmphiF$Diel <- as.factor(AmphiF$Diel)
```

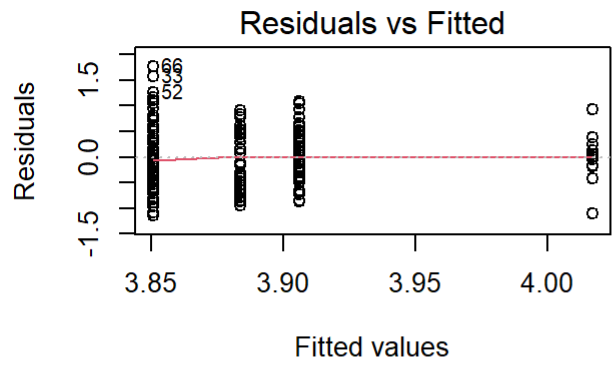
```
par(mfrow=c(2,2), mar = c(3,5,4,5))

#Histogram
line = 2.5
cex = 1
side = 3
adj=-0.44

#Body Size
hist(AmphiF$Body_size_mm, xlab="Body Size(mm)", main = "Amphibian Body Size (n=200)", col=rgb(15
3,255,255,max=255))
mtext("A", side=side, line=line, cex=cex, adj=adj)

#Reproductive Output
hist(AmphiF$Reproductive_output_y, xlab="Reproductive Output", main = "Amphibian Reproductive Ou
tput (n=200)", col=rgb(100,150,120,max=255))
mtext("B", side=side, line=line, cex=cex, adj=adj)

#Log Transformation Required
#Body Size
AmphiF$log_amphi <- log(AmphiF$Body_size_mm+1)
hist(AmphiF$log_amphi, xlab="Log Body Size(mm)", main = "Log Amphibian Body Size (n=200)", col=r
gb(58,87,149,max=255))
mtext("C", side=side, line=line, cex=cex, adj=adj)

#Reproductive Output
AmphiF$log_rep <- log(AmphiF$Reproductive_output_y+1)
hist(AmphiF$log_rep, xlab="Log Reproductive Output", main = "Log Amphibian Reproductive Output
 (n=200)", col=rgb(200,120,120,max=255))
mtext("D", side=side, line=line, cex=cex, adj=adj)
```

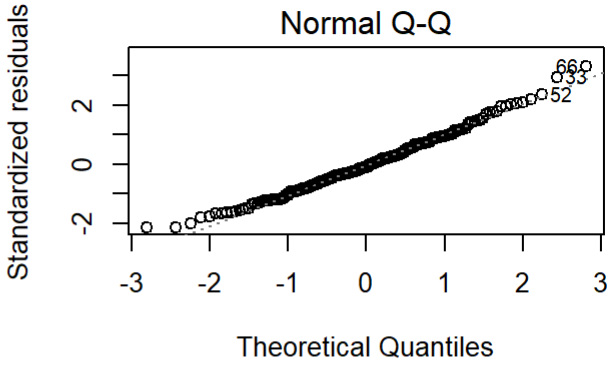


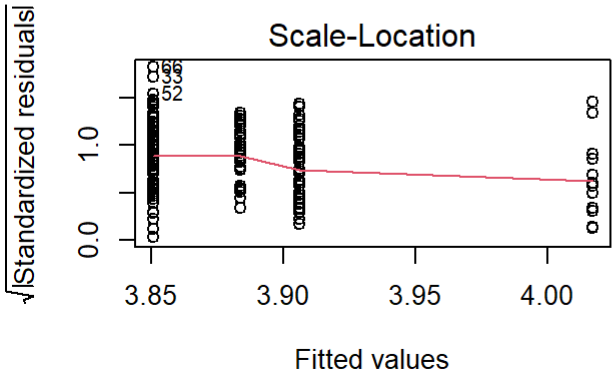

```
#Box plot
p <- ggplot(AmphiF, aes(x= Habitat, y=log_amphi,fill=Habitat))
bxp <- p + geom_boxplot() +
labs(x="Habitat", y="log Body Size (mm)") +
scale_fill_discrete(labels=c("Aquatic","Arboreal","Fossorial","Terrestrial"))

q <- ggplot(AmphiF, aes(x= Diel, y=log_amphi,fill=Diel))
bxp1 <- q + geom_boxplot() +
labs(x="Diel Activity", y="log Body Size (mm)") +
scale_fill_discrete(name= "Diel Activity",labels=c("Crepuscular","Diurnal","Nocturnal"))

figure <- ggarrange(bxp, bxp1,
                    labels = c("A", "B"),
                    ncol = 1, nrow = 2)
figure
```

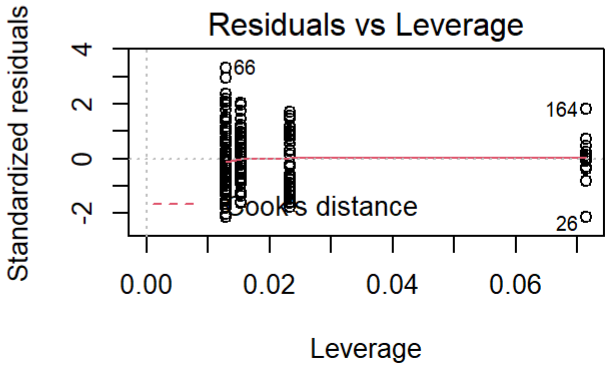



```
#Habitat Regression Plot
fit_amphi <- lm(log_amphi~Habitat, data=AmphiF)
par(mfrow=c(2,2))
plot(fit_amphi)
```

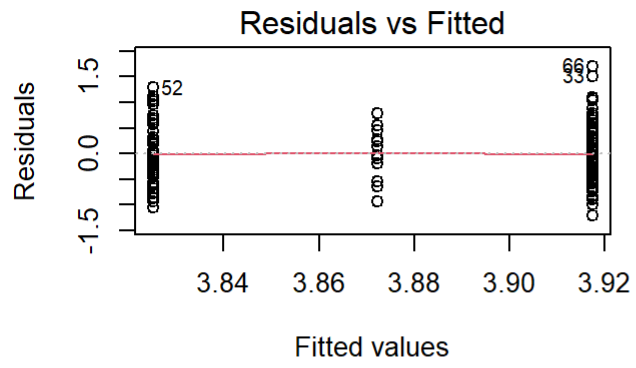


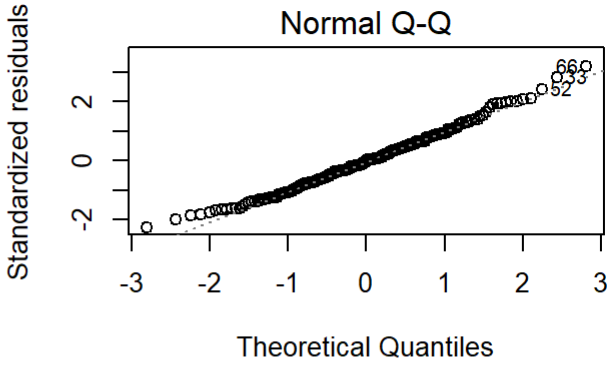


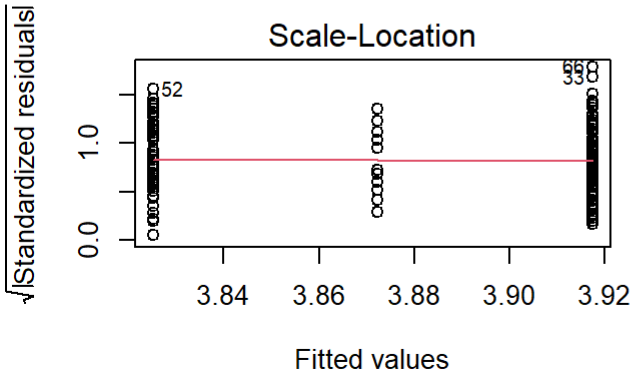


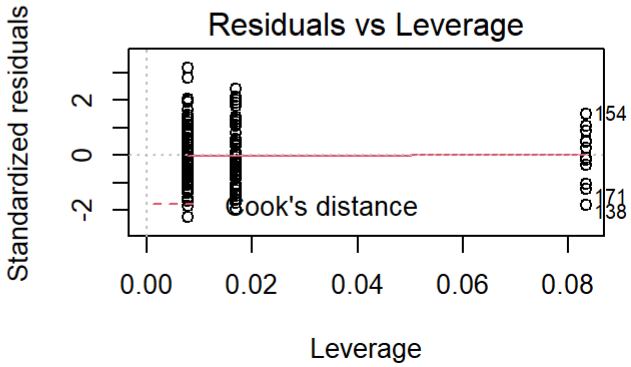



```
#Diel Regression Plot
fit_amphi2 <- lm(log_amphi~Diel, data=AmphiF)
par(mfrow=c(2,2))
plot(fit_amphi2)
```

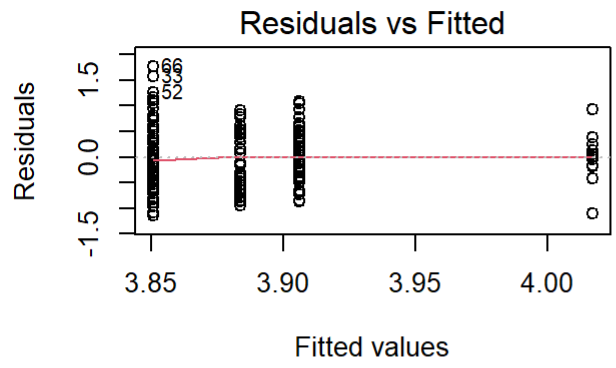


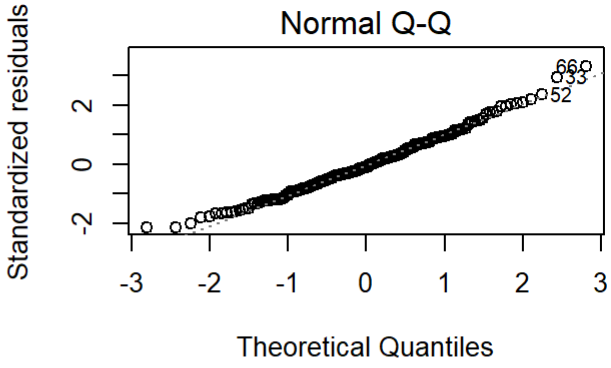


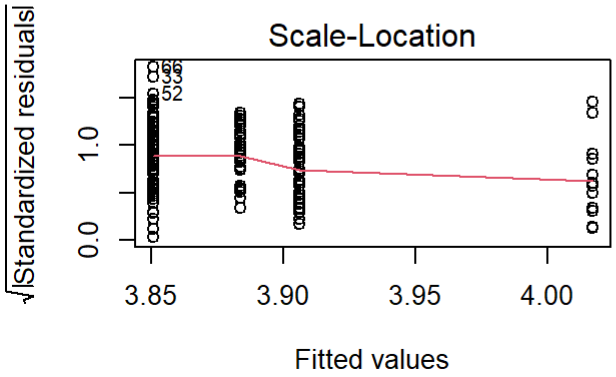


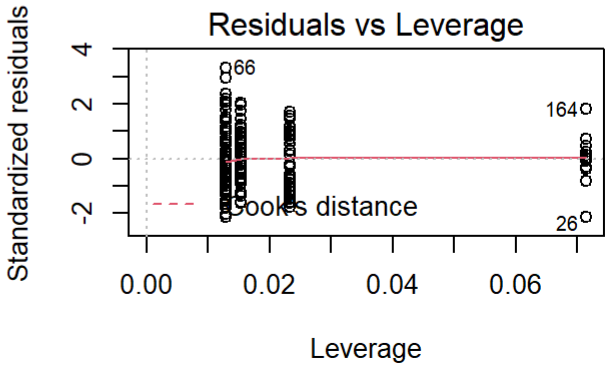



```
#ANOVA Test (Habitat)
fit_amphi <- lm(log_amphi~Habitat, data=AmphiF)
par(mfrow=c(2,2))
plot(fit_amphi)
```





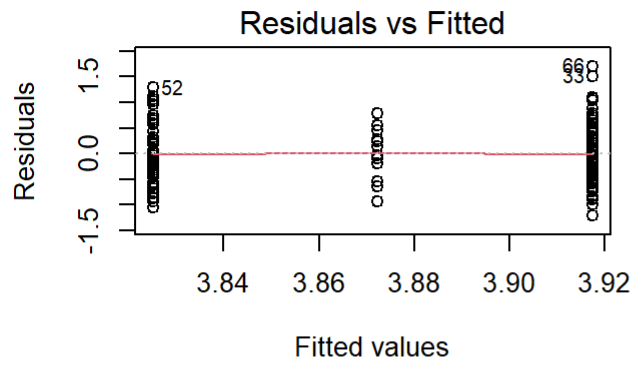


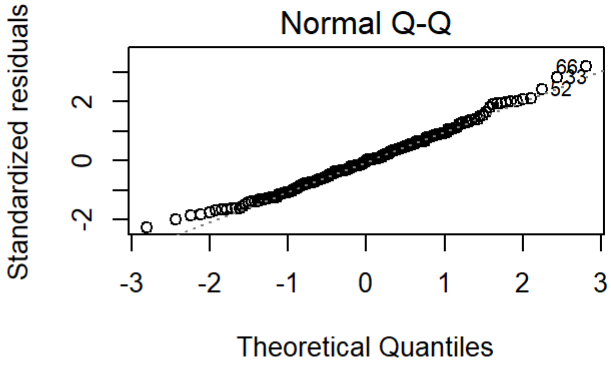


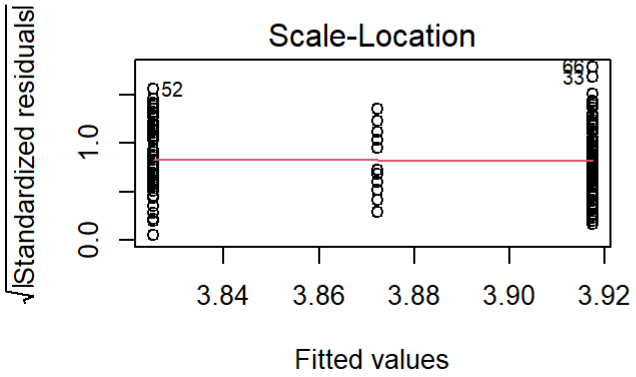

```
par(mfrow=c(1,1))
aov_amphi <- aov(log_amphi ~ Habitat, data=AmphiF)
summary(aov_amphi)
```

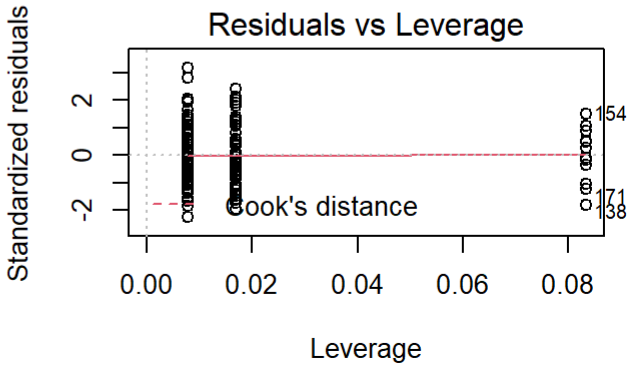
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Habitat       3   0.36  0.1216   0.421  0.738
## Residuals   196  56.67  0.2891
```

```
#Anova Test(Diel)
#Diel
fit_amphi2 <- lm(log_amphi~Diel, data=AmphiF)
par(mfrow=c(2,2))
plot(fit_amphi2)
```





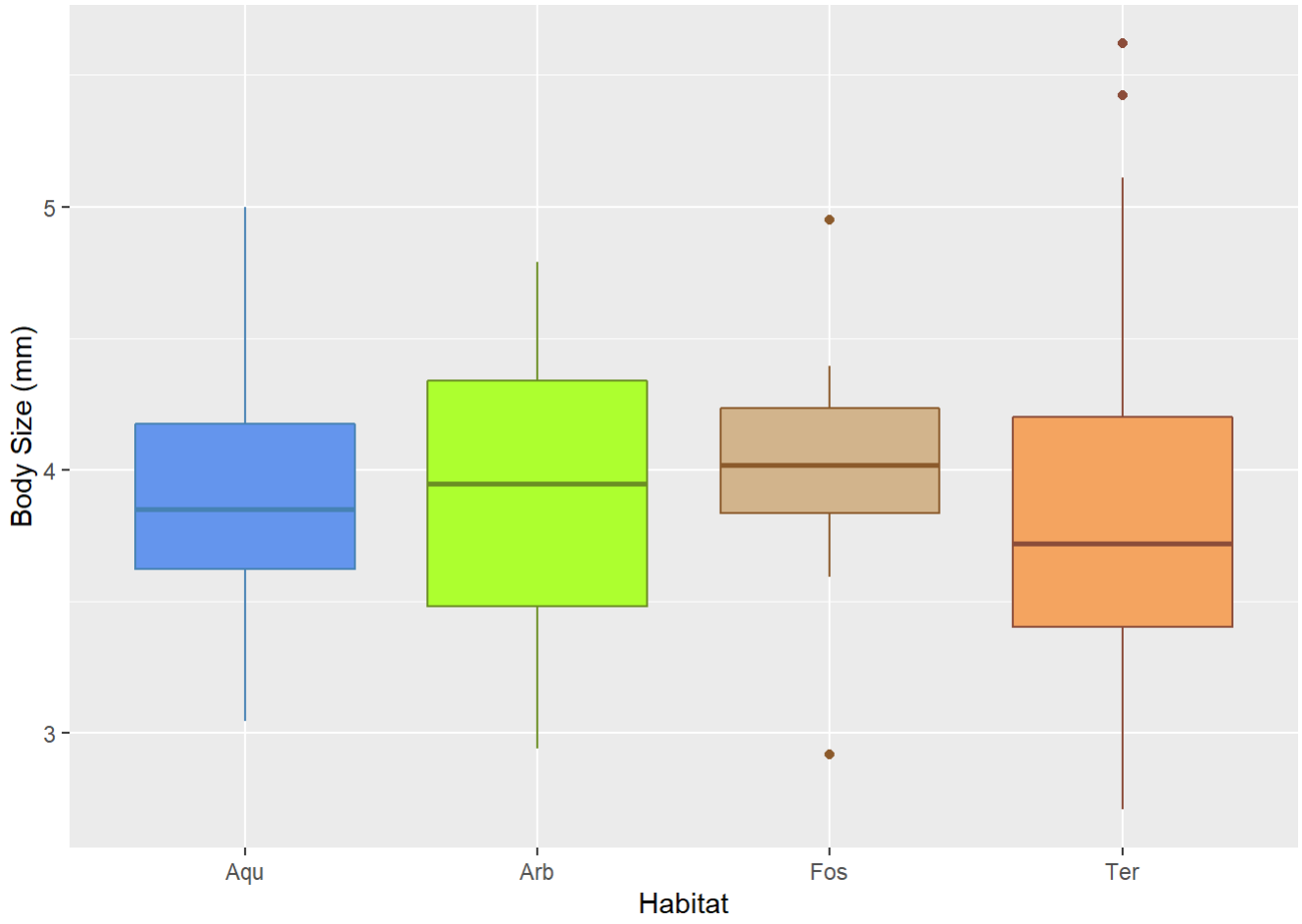




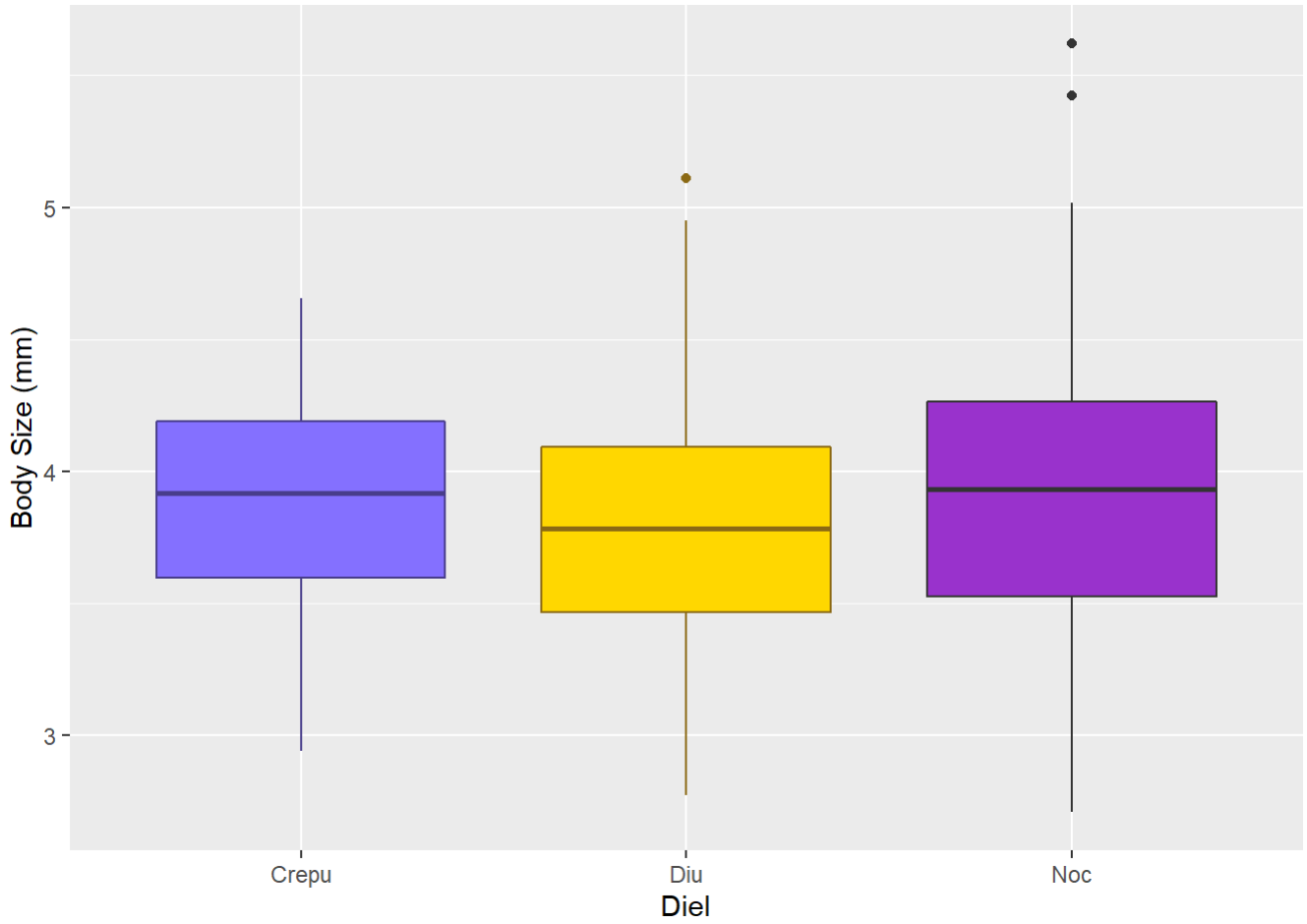

```
par(mfrow=c(1,1))
aov_amphi2 <- aov(log_amphi ~ Diel, data=AmphiF)
summary(aov_amphi2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Diel          2   0.35  0.1741   0.605  0.547
## Residuals   197  56.68  0.2877
```

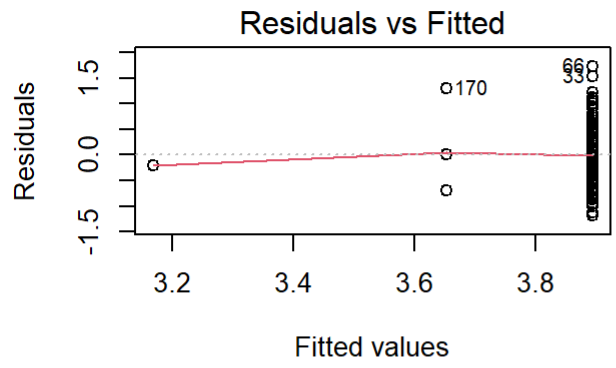
```
#Anova Visualization (Habitat)
ggplot(AmphiF, aes(x=Habitat, y=log_amphi, fill=Habitat, color=Habitat))+
ylab("Body Size (mm)")+
geom_boxplot(show.legend=F)+ #do a boxplot, but hide legend cause its just colors
scale_fill_manual(values=c("cornflowerblue","greenyellow","tan","sandybrown"))+
scale_color_manual(values=c("steelblue", "olivedrab", "tan4","salmon4"))
```

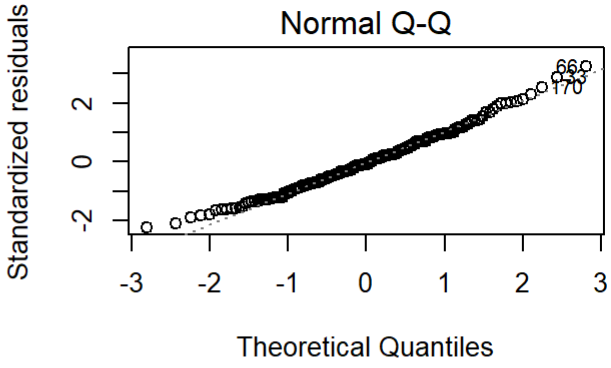


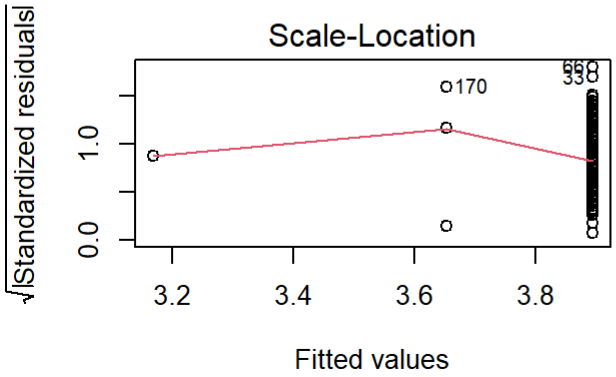

```
#Anova Visualization (Diel)
ggplot(AmphiF, aes(x=Diel, y=log_amphi, fill=Diel, color=Diel))+ #base plot
geom_boxplot(show.legend=F)+
ylab("Body Size (mm)")+
scale_fill_manual(values=c("lightslateblue","gold","darkorchid"))+ #set the fill colors
scale_color_manual(values=c("darkslateblue", "goldenrod4", "grey20"))
```

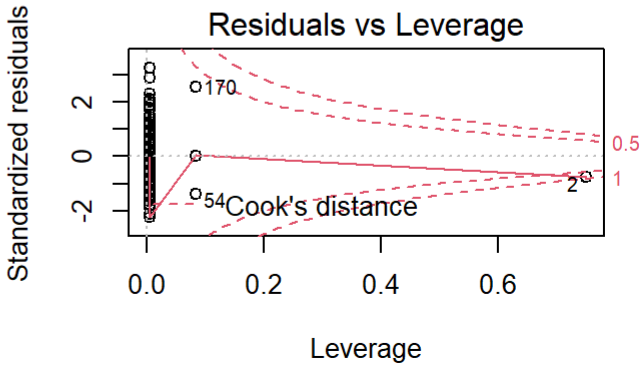



```
#Linear Regression
amp.reg <- lm(log_amphi ~ Reproductive_output_y, data=AmphiF)
par(mfrow=c(2,2))
plot(amp.reg)
```










```
summary(amp.reg)
```

```
##
## Call:
## lm(formula = log_amphi ~ Reproductive_output_y, data = AmphiF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18674 -0.37878 -0.04465  0.36789  1.72561
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.1369     0.1642  25.195   <2e-16 ***
## Reproductive_output_y  -0.2421     0.1552  -1.561     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5334 on 198 degrees of freedom
## Multiple R-squared:  0.01215,    Adjusted R-squared:  0.007163
## F-statistic: 2.436 on 1 and 198 DF,  p-value: 0.1202
```

```
#Linear Regression Visualization

ggplot(AmphiF, aes(x=Reproductive_output_y, y=log_amphi))+
geom_jitter(size=3, alpha=0.6, shape=21)+
geom_smooth(method="lm", alpha=0.6, size=0.5)+
scale_color_manual(values=c("orchid"))+
scale_fill_viridis_d("orchid")+
ylab("Body Size (mm)")+
xlab("Reproductive Output")
```

```
## `geom_smooth()` using formula 'y ~ x'
```