EEMB 146 Final Project

Samantha Chen

5/24/2021

Abstract

In this R analysis, global crayfish data from an article by Bland (2017) was used to determine if body size is affected by external factors such as family, extinction risk, habitat type, and human population density. Results of this study showed variation in body size between families and a risk of decreasing body size with increased human population density. More research will be needed to explore all other factors as this study only scratches the surface of human impacts on crayfish.

Introduction

Freshwater crayfish does not only play an important role in aquatic food webs, but are also of great economic importance as one of the most sought-after seafood (Jones et. al, 2006). However, with the worsening of climate change and overfishing, crayfish abundance is declining, threatening both the integrity of our environment as well as food security in vulnerable countries. In a global comparative study on crayfish extinction risk, Bland (2017) discovered significant relationships between crayfish characteristics (e.g. chela shape, egg number, etc.) and susceptibility to extinction (IUCN Red List). To further add onto Bland's study, this statistical analysis will be conducted using the same dataset to look at how body size of crayfish is affected by various explanatory variables including Family, Human Popultion Density (HPD), Habitat Type, and IUCN rating. The two questions that this analysis will be asking is (1) is there a difference in mean body size between the different families of crayfish and (2) are any of the variables of interest a good predictor of body size?

Exploratory Data Analysis

Preliminary data analysis was conducted to visualize and check if the assumptions of normality and homogeneity of variances are met. The response variable, body size, measured in mm, was not normally distributed even after log and square root transformations (*Figure 1*). The only continuous predictor is human population density (HPD) measured in people per km^2. As for the categorical predictors, Family had three levels (Astacidae, Cambaridae, Parastacidae) but Astacidae was omitted due to how small its sample size was compared to that of the other two families. Habitat type had four levels (Burrows, Caves, Lakes/Wetland, Streams/River) and IUCN rating had five levels (Least Concern, Near Threatened, Vulnerable, Endangered, Critical). A boxplot of Family showed that Parastacidae had more in-between variation than does Cambaridae (*Appendix 1*). Visualizations of all other predictors showed similar medians with Habitat having the most in-between variation (*Appendix 2*). There was no colinearity between predictors of interest and body size had largest correlation with habitat type (*Figure 2*).

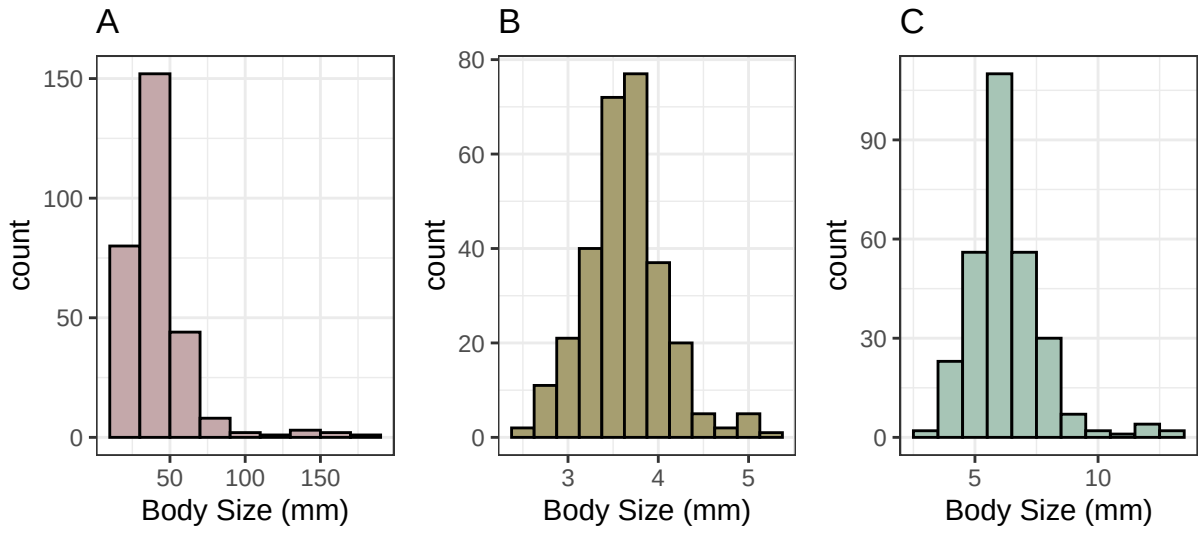


Figure 1: (A) Histogram of untransformed crayfish body size showing a right skewed distribution. (B) Histogram of log transformed crayish body size with bell shape curve but not true normality (shapiro.wilk test $p<0.05$). (C) Histogram of square root transformed crayfish body size with right skewed distribution.

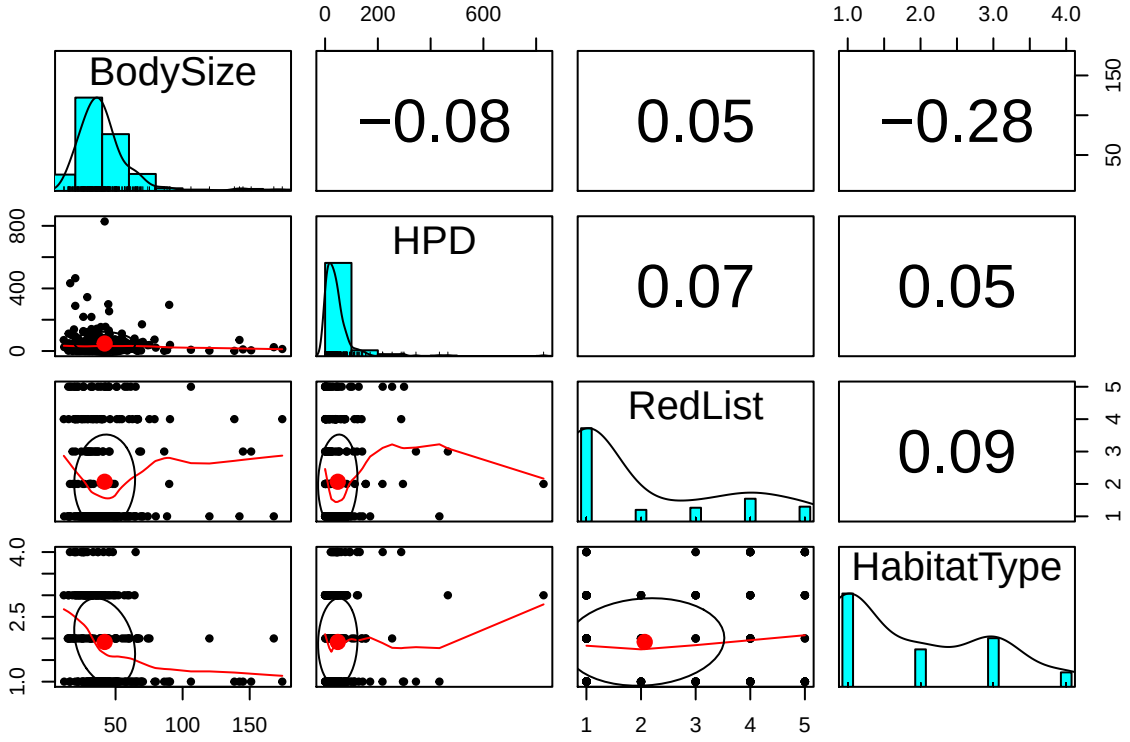


Figure 2: Correlation plot between predictors of interest and body size. There is no sign of colinearity (R>0.6) and therefore no strong interaction between predictors. Body size and Habitat type show the strongest correlation (R = -0.28) out of all predictors.

Statistical Methods

A two-sample t-test was used to determine if there was a significant difference in average body size between different families, but because the residuals of body size was not normal and the two families had unequal variances (see appendix) a non-parametric t-test was used instead. This was the same case when analyzing the relationship between body size and predictors of interest.

Do crayfish body size vary by Family?

A Wilcoxon Rank Sum Test was used to test if the two families had the same shape. The null hypothesis is that there is no difference in shape and therefore have the same body size median. The alternative hypothesis is that there is a difference in shape, therefore they have different body size medians.

What variables of interest are good predictors of crayfish body size?

A Poisson regression was used to test which of the three explanatory variables of interest (HPD, Habitat, IUCN rating) is a good predictor of crayfish body size.

Results

Do crayfish body size vary by Family?

The Wilcoxon Rank Sum Test produced a p-value less than alpha = 0.05 (*see Appendix 3*), therefore the null hypothesis that there is no shape difference between the two families is rejected. There is a difference between the shapes of the two families and median body size between families is different.

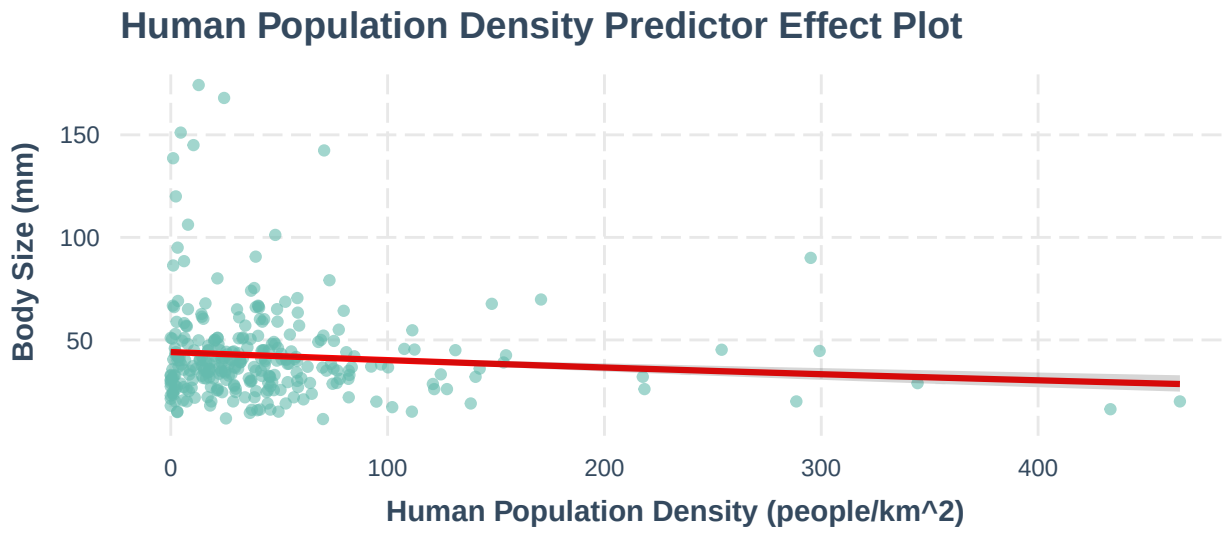


Figure 3: Effect plot of HPD as a predictor of body size. The intercept of the model is 43.115 (people/kmˆ2) and the slope is 0.9994 (p<0.0001). There is a negative relationship between HPD and body size.

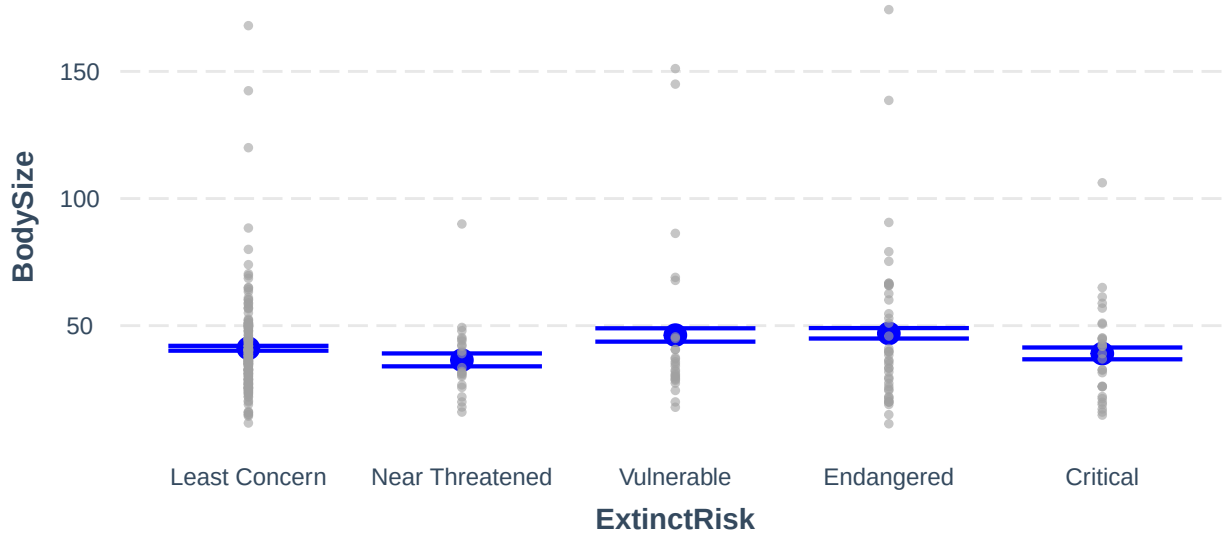


Figure 4: Effect plot of extinction risk as a predictor of body size. Each rating is compared against the control, Least Concern, which contains the least amount of variation. Vulnerable (p=0.000143), Near threatened (p=0.001265), and Endangered (p<0.0001) are significant coefficients in the model while Critical is not (p=0.110358).

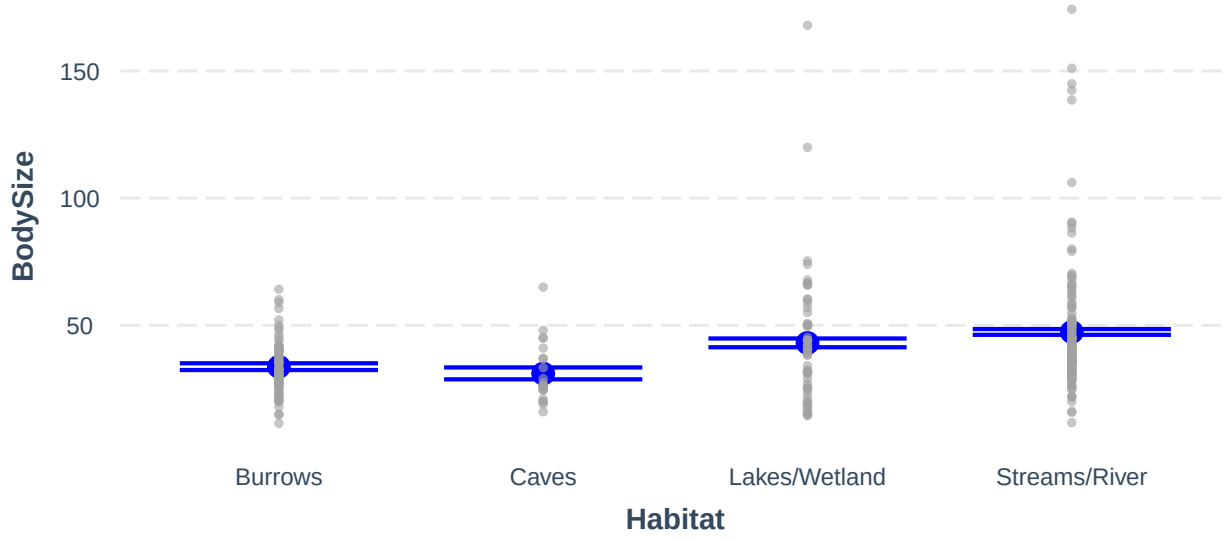


Figure 5: Effect plot of habitat type as a predictor of body size. Each habitat type is compared against Burrows, containing the least amount of variation. Cave is not a significant coefficient in the model (p=0.0524), output=FALSE, while Lakes/Wetland and Streams/River are highly significant coefficients of the model (p<0.0001). Most in between variation in body size is found in Caves.

What variables of interest are good predictors of crayfish body size?

Human Population Density was a highly significant predictor of crayfish body size ($p<0.05$) and the poisson regression equation is body size = 0.9994(HPD) + 43.115 (*see Appendix 4 for calculations*). The effect plot shows a negative relationship in which body size decreases with human population density (*Figure 3*). IUCN Red List rating and Habitat are also significant predictors of crayfish body size ($p<0.05$). For IUCN Rating, Near Threatened, Vulnerable, and Endangered are significantly different from the control, Least Concern (*Figure 4*). For Habitat, there was no true control but other habitats were compared against Burrows. Lakes/Wetlands and Streams/River were found to be significantly different from Burrows (*Figure 5*).

Discussion

From this analysis there is evidence that crayfish body size vary with family and that, by looking at the boxplot in *Appendix 1*, crayfish in the Parastacidae family tend to be larger than crayfish in the Cambaridae family. However, an ad-hoc test will be needed to determine the specific differences in body size. For the regression analysis, while all three predictors of interest had significant results, the valuable predictors are human population density and habitat type. As human population density increased, body size decreased, which is a familiar pattern in studies on anthropogenic impacts on wildlife. While habitat type and IUCN rating had similar results, Habitat was the better predictor because it was significantly correlated with body size.

Much of the limitations of this study is that the data did not meet many assumptions that would have allowed for stronger tests. Despite that, this study is a stepping stone towards using R as a tool for meta-analysis. The results of this study is preliminary data but it warns of the negative impacts human have on freshwater crayfish. More in-depth analysis with larger datasets, as well as more variables, will be needed to determine the future of these crustaceans.

References

Journal Articles

Jones, J., Andriahajaina, F., Ranambinintsoa, E., Hockley, N., & Ravoahangimalala, O. (2006). The economic importance of freshwater crayfish harvesting in Madagascar and the potential of community-based conservation to improve management. Oryx, 40(2), 168-175. doi:10.1017/S0030605306000500

Bland, L. (2017). Global correlates of extinction risk in freshwater crayfish. Animal Conservation, 20(6), 532-542.

Packages

Hadley Wickham and Jim Hester (2020). readr: Read Rectangular Text Data. R package version 1.4.0. https://CRAN.R-project.org/package=readr

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.5. https://CRAN.R-project.org/package=dplyr

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

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/

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Tim Bergsma (2018). latexpdf: Convert Tables to PDF or PNG. R package version 0.1.6. https://CRAN.R-project.org/package=latexpdf

Long JA (2020). jtools: Analysis and Presentation of Social Scientific Data. *R package version 2.1.0, <URL: https://cran.r-project.org/package=jtools>.*

Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2.https://CRAN.R-project.org/package=RColorBrewer

Revelle, W. (2020) psych: Procedures for Personality and Psychological Research, Northwestern,University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.1.3,.

Thomas Lin Pedersen (2020). patchwork: The Composer of Plots. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork

Appendix

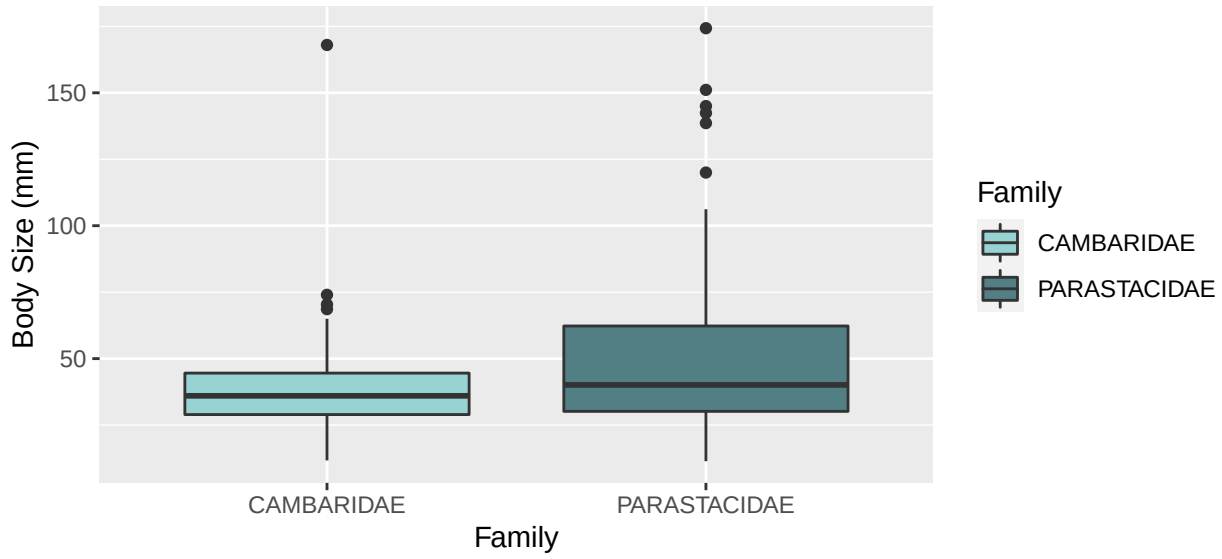
Data Preparation

```
# load and subset
dataset_6 <- read.csv("~/EEMB 146 Lab Files/Final Project/dataset_6.xlsx_csv.csv")
# dim(dataset_6) --> 300 observations, 26 variables
dataset_sub <- dataset_6 %>%
  select(Family, RedList, BodySize, HabitatType, HPD)
# turning body size into numeric variable
dataset_sub$BodySize <- as.numeric(as.character(dataset_sub$BodySize))
#  subsetting out Astacidae Family
dataset_cleaned <- dataset_sub[!(dataset_sub$Family == "ASTACIDAE"),]
# subsetting data for comparing body size means
family_bodysize <- dataset_cleaned %>%
  select(Family, BodySize)
cambaridae_data <- subset(dataset_cleaned, Family == "CAMBARIDAE")
parastacidae_data <- subset(dataset_cleaned, Family == "PARASTACIDAE")
# changing RedList into categorical variable
dataset_cleaned$RedList <- as.factor(dataset_cleaned$RedList)
dataset_cleaned$ExtinctRisk[dataset_cleaned$RedList == "1"] <- "Least Concern"
dataset_cleaned$ExtinctRisk[dataset_cleaned$RedList == "2"] <- "Near Threatened"
dataset_cleaned$ExtinctRisk[dataset_cleaned$RedList == "3"] <- "Vulnerable"
dataset_cleaned$ExtinctRisk[dataset_cleaned$RedList == "4"] <- "Endangered"
dataset_cleaned$ExtinctRisk[dataset_cleaned$RedList == "5"] <- "Critical"
#changing Habitat into categorical variable
dataset_cleaned$Habitat[dataset_cleaned$HabitatType == "1"] <- "Streams/River"
dataset_cleaned$Habitat[dataset_cleaned$HabitatType == "2"] <- "Lakes/Wetland"
dataset_cleaned$Habitat[dataset_cleaned$HabitatType == "3"] <- "Burrows"
dataset_cleaned$Habitat[dataset_cleaned$HabitatType == "4"] <- "Caves"
# log transform body size
dataset_cleaned$log_BodySize <- log(dataset_cleaned$BodySize)
# sqrt transforming bodysize
dataset_cleaned$sqrt_BodySize <- sqrt(dataset_cleaned$BodySize)
# Subsetting and adjusting dataset for poisson regression
dataset_poisson <- dataset_cleaned %>%
  select(BodySize, HPD, RedList, HabitatType)
# subsetting ExtinctRisk for poisson regression
extinctrisk_bodysize <- dataset_cleaned %>%
```

```
  select(ExtinctRisk, BodySize)
# subsetting Habitat for poisson regression
habitat_bodysize <- dataset_cleaned %>%
  select(Habitat, BodySize)
```

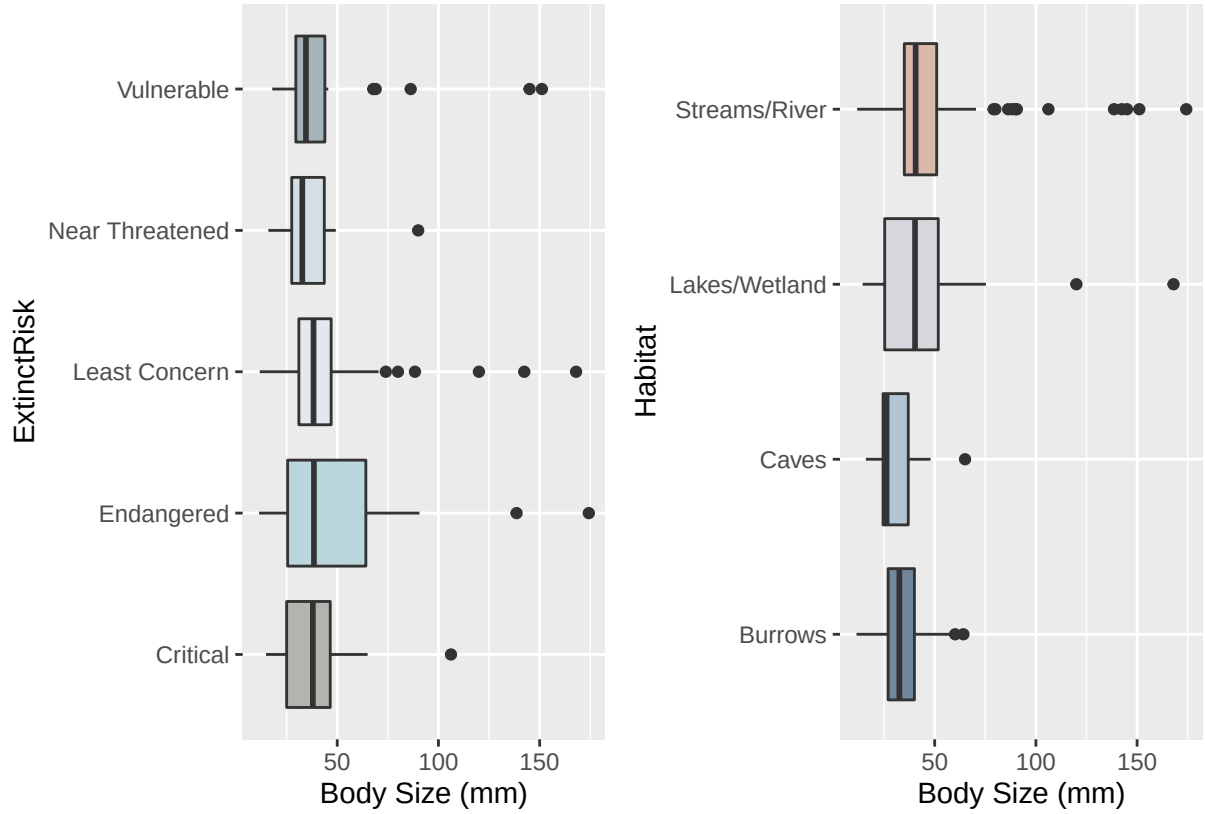
Appendix 1

```
# ggplot boxplot for family
Family_boxplot <- ggplot(family_bodysize, aes(x = Family, y = BodySize, fill = Family)) +
  geom_boxplot() +
  ylab("Body Size (mm)")
Family_boxplot + scale_fill_manual(values = c("#98d3d4", "#517f83"))
```



Appendix 2

```
# ggplot boxplot for Extinct Risk
Redlist_boxplot <- ggplot(dataset_cleaned, aes(x = ExtinctRisk, y = BodySize,
                                               fill = ExtinctRisk)) +
  geom_boxplot() +
  scale_fill_manual(values = c("#b4b4af", "#bad5dc", "#e3e8ee", "#d4dfe5", "#a5b4b9")) +
  theme(legend.position = "none") +
  ylab("Body Size (mm)")
# ggplot boxplot for habitat type
Habitat_boxplot <- ggplot(dataset_cleaned, aes(x = Habitat, y = BodySize, fill = Habitat)) +
  geom_boxplot() +
  scale_fill_manual(values = c("#71879a", "#b1c5d4", "#d6d7dc", "#d8b9aa")) +
  ylab("Body Size (mm)") +
  theme(legend.position = "none")

Redlist_boxplot + coord_flip() + Habitat_boxplot + coord_flip()
```



Testing Normality and Homogeneity of Variances

```
par(mfrow = c(2,2))
# Testing normality of log transformed data
qqPlot(dataset_cleaned$log_BodySize)
```

```
## [1] 258 234
```

```
shapiro.test(dataset_cleaned$log_BodySize)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset_cleaned$log_BodySize
## W = 0.9759, p-value = 7.621e-05
```

```
# p-value = 7.621e-05
# we reject null hypothesis that response variable is normal

# Testing normality of sqrt transformed data
qqPlot(dataset_cleaned$log_BodySize)
```

```
## [1] 258 234
```

```
shapiro.test(dataset_cleaned$log_BodySize)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset_cleaned$log_BodySize
## W = 0.9759, p-value = 7.621e-05
```

```
# p-value = 7.621e-05
# we reject null hypothesis that response variable is normal

# Testing homogeneity of variances for family
leveneTest(dataset_cleaned$BodySize, dataset_cleaned$Family)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1  23.853 1.718e-06 ***
##       291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

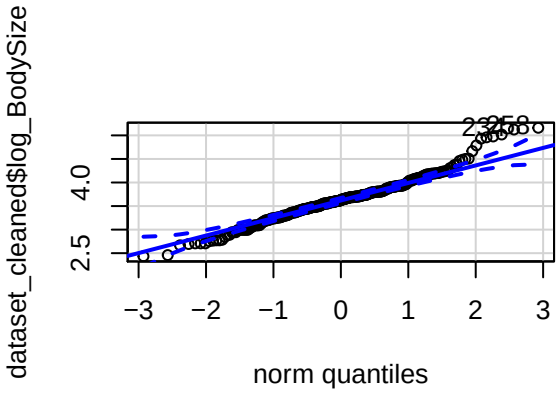
```
# p-value = 1.718e-06
# reject null that the variances are equal

#Testing normality of body size residuals
fit_family <- lm(BodySize ~ Family, data = dataset_cleaned)
res_family = fit_family$residuals

# testing residuals for normality
shapiro.test(res_family)
```

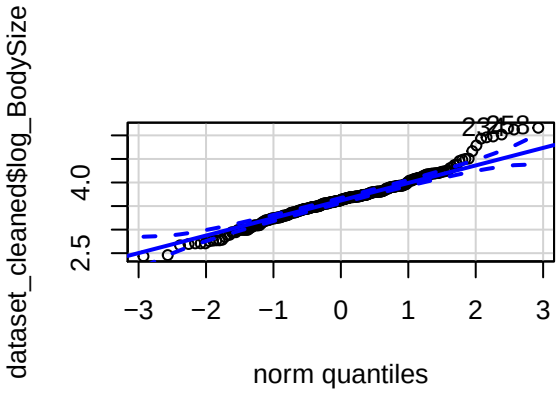
```
##
##  Shapiro-Wilk normality test
##
## data:  res_family
## W = 0.79942, p-value < 2.2e-16
```

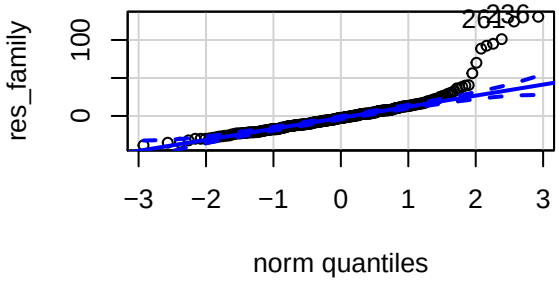
```
# p-value < 2.2e-16
# reject the null that the residuals are normal
qqPlot(res_family)
```

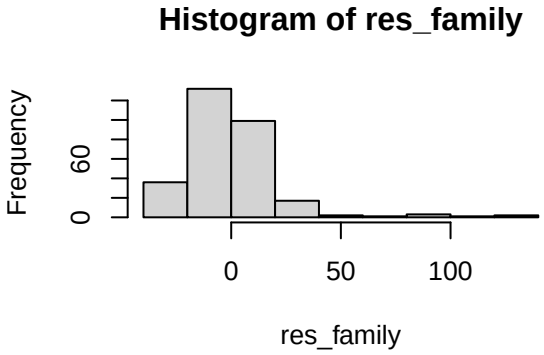
```
## 236 261
## 233 257
```

```
hist(res_family)
```










```
# Distribution is right skewed which is consistent with qqplot and shapiro wilk test
```

Appendix 3

```
wilcox.test(cambaridae_data$BodySize, parastacidae_data$BodySize) # p-value = 0.002496
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  cambaridae_data$BodySize and parastacidae_data$BodySize
## W = 7651.5, p-value = 0.002496
## alternative hypothesis: true location shift is not equal to 0
```

```
#null = true location shift is equal to 0
#alt = true location shift not = 0
```

Appendix 4

```
# Will be doing a poisson glm for HPD
mod.crayfish_HPD <- glm(BodySize ~ HPD, family = poisson, data = dataset_poisson)
summary(mod.crayfish_HPD)
```

```
##
## Call:
```

```
## glm(formula = BodySize ~ HPD, family = poisson, data = dataset_poisson)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5954  -2.0115  -0.6437   0.9341  15.0577
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.7638701  0.0109280 344.426  < 2e-16 ***
## HPD         -0.0006335  0.0001370  -4.624 3.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2822.0  on 292  degrees of freedom
## Residual deviance: 2798.7  on 291  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

```
# plotting
effect_plot(mod.crayfish_HPD, glm = TRUE, pred = HPD, interval = TRUE, plot.points = TRUE)
```

```
# one large outlier is affecting fit of glm --> will need to subset out
# outlier = 828.15976330

# plotting HPD regression without outlier
mod.crayfish_HPD2 <- glm(BodySize ~ HPD, family = poisson, data = dataset_poisson2)
effect_plot(mod.crayfish_HPD2, glm = TRUE, pred = HPD, interval = TRUE, plot.points = TRUE,
            main.title = "Human Population Density Predictor Effect Plot",
            x.label = "Human Population Density (people/km^2)",
            y.label = "Body Size (mm)",
            point.color = "#65bbae",
            colors = "red",
            line.thickness = 0.75) # negative relationship
```

```
# finding slope and intercept
HPD_slope <- exp(-0.0006335)
HPD_intercept <- exp(3.7638701)
print(HPD_slope) #slope = 0.9994 (approximately 1)
```

```
## [1] 0.9993667
```

```
print(HPD_intercept) # intercept =43.115
```

```
## [1] 43.11496
```

Poisson Regression for IUCN Rating

```r
extinctrisk_bodysize$ExtinctRisk <- factor(extinctrisk_bodysize$ExtinctRisk,
                                           levels = c("Least Concern", "Near Threatened", "Vulnerable",

# poisson for RedList
mod.crayfish_ExtinctRisk <- glm(BodySize ~ ExtinctRisk, family = poisson,
                                data = extinctrisk_bodysize)
summary(mod.crayfish_ExtinctRisk)
```

```
##
## Call:
## glm(formula = BodySize ~ ExtinctRisk, family = poisson, data = extinctrisk_bodysize)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.2310  -2.1688  -0.6984   1.0489  14.8090
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.71588    0.01183 314.217  < 2e-16 ***
## ExtinctRiskNear Threatened -0.12007    0.03724  -3.224 0.001265 **
## ExtinctRiskVulnerable       0.11851    0.03116   3.803 0.000143 ***
## ExtinctRiskEndangered       0.13318    0.02520   5.284 1.26e-07 ***
## ExtinctRiskCritical        -0.05186    0.03248  -1.597 0.110358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2822.0  on 292  degrees of freedom
## Residual deviance: 2760.6  on 288  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: Inf
##
## Number of Fisher Scoring iterations: 5
```

```r
#plotting
effect_plot(mod.crayfish_ExtinctRisk, pred = ExtinctRisk, interval = TRUE,
            plot.points = TRUE,
            point.color = "#a1a1a1",
            colors = "blue",
            point.size = 1,
            line.thickness = 0.75)
```

```r
# Not a lot of variation, coefficient very small
#Vulnerable has a little more variation (looking at error bars)
# Like an ANOVA but comparing against Poisson Distribution
#currently comparing other categories against critical
# wouldn't need AIC because I am only using one predictor for each model
# exp() anything for poisson distribution
# Near threatened, Vulnerable, and Endangered significantly different from Least Concern
```

Poisson Regression for Habitat

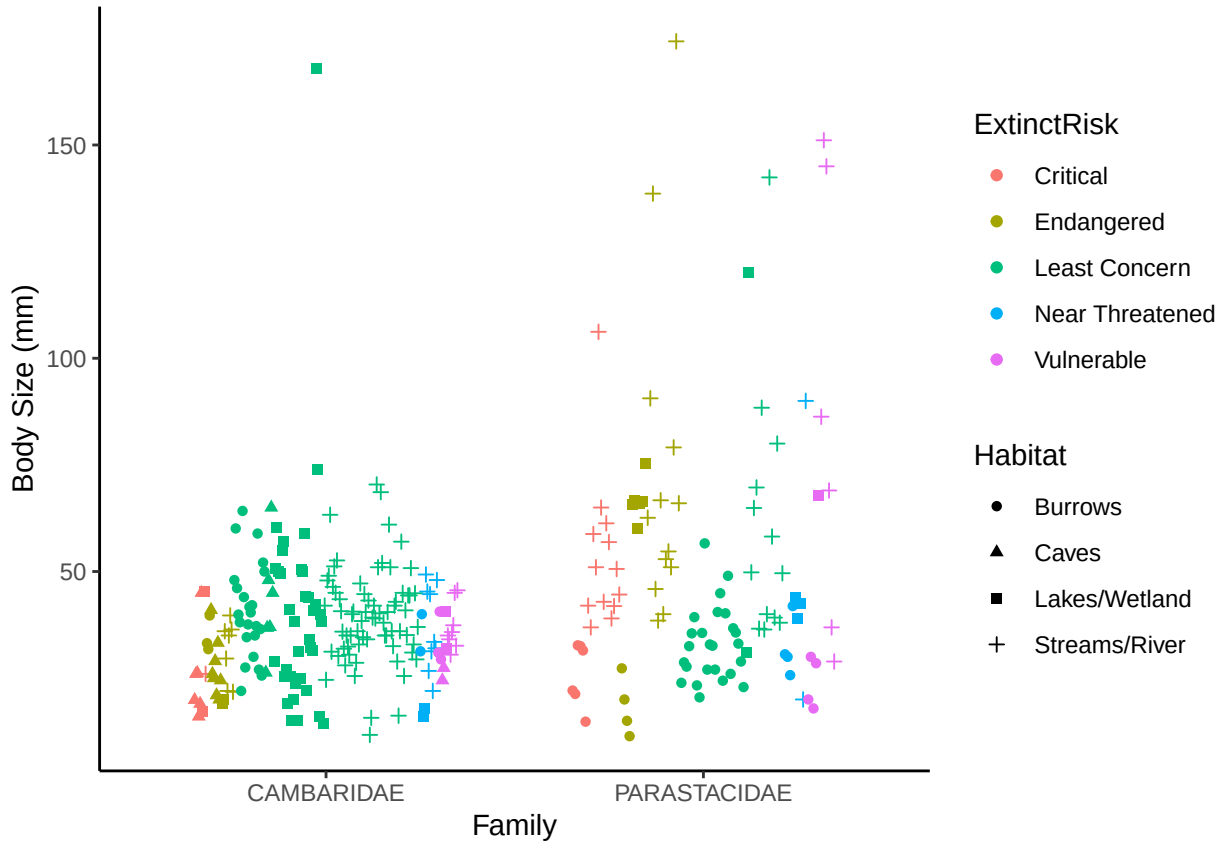
```r
# poisson for RedList
mod.crayfish_Habitat <- glm(BodySize ~ Habitat, family = poisson, data = habitat_bodysize)
summary(mod.crayfish_Habitat)
```

```
##
## Call:
## glm(formula = BodySize ~ Habitat, family = poisson, data = habitat_bodysize)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.2145  -1.8857  -0.6505   1.0124  14.4075
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            3.51766    0.02002 175.679   <2e-16 ***
## HabitatCaves          -0.08382    0.04321  -1.940   0.0524 .
## HabitatLakes/Wetland   0.24479    0.02856   8.571   <2e-16 ***
## HabitatStreams/River   0.34037    0.02347  14.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2822.0  on 292  degrees of freedom
## Residual deviance: 2527.7  on 289  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

```r
#plotting
effect_plot(mod.crayfish_Habitat, pred = Habitat, interval = TRUE,
            plot.points = TRUE,
            point.color = "#a1a1a1",
            colors = "blue",
            point.size = 1,
            line.thickness = 0.75)
```

Fun Extra Plot

```r
# visualizing all categorical data
ggplot(data = dataset_cleaned,
       aes(x = Family, y = BodySize, fill = ExtinctRisk, color = ExtinctRisk,
           shape = Habitat)) +
  geom_jitter(size = 1.5, position = position_dodge2(width = 0.7)) +
  labs(x = "Family", y = "Body Size (mm)") +
  theme_classic()
```




```
# Parastacidae more variation than Cambaridae
# Distinct grouping of body size with habitat groups and extinction risks
```