Chapter 5 Homework Part B

David Oniani

February 19, 2021

```
# Import libraries
library(MASS)
library(Stat2Data)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

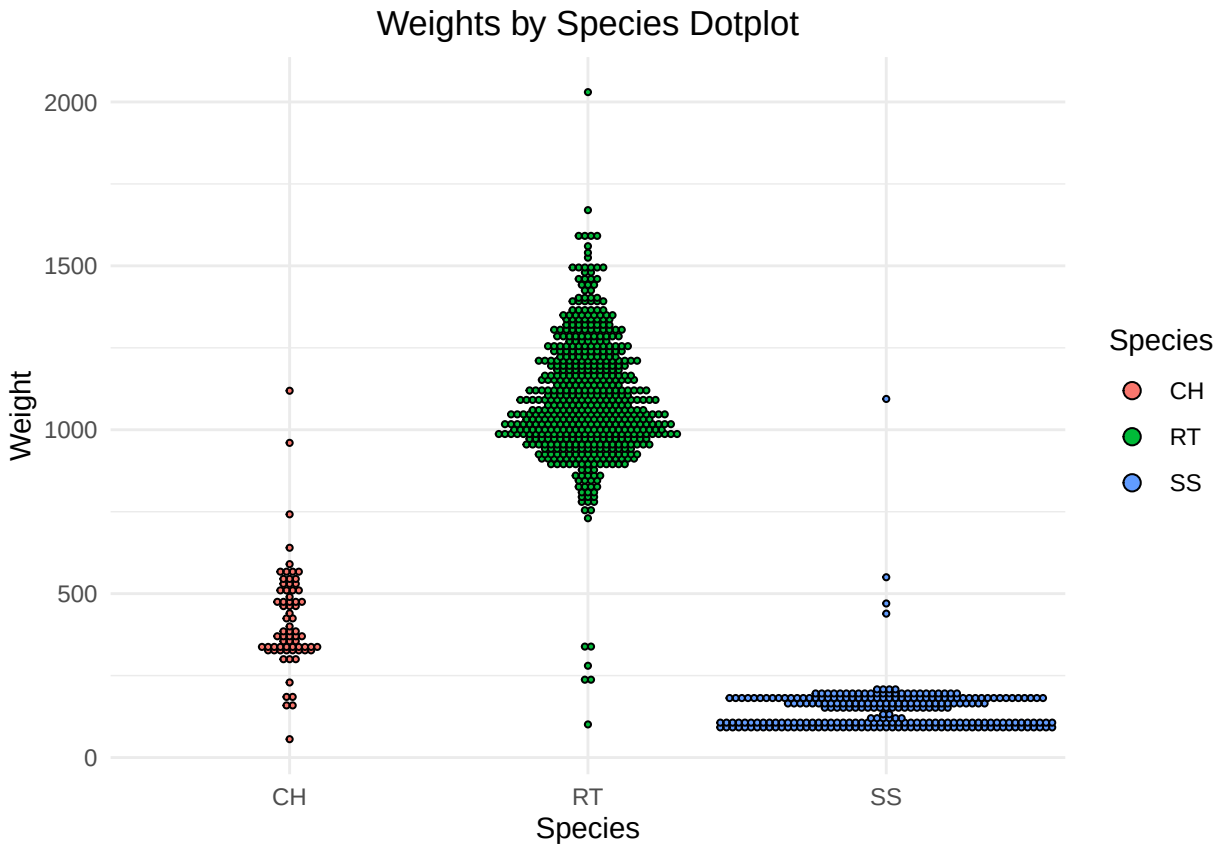
```
library(emmeans)
library(ggplot2)  # Import the coolest library of R
```

Exercise 5.32

Do Exercise 5.32 as stated. Also, find a suitable response transformation and and do a one-way ANOVA on that transformed scale. Include residual analysis and all pairwise group comparisons.

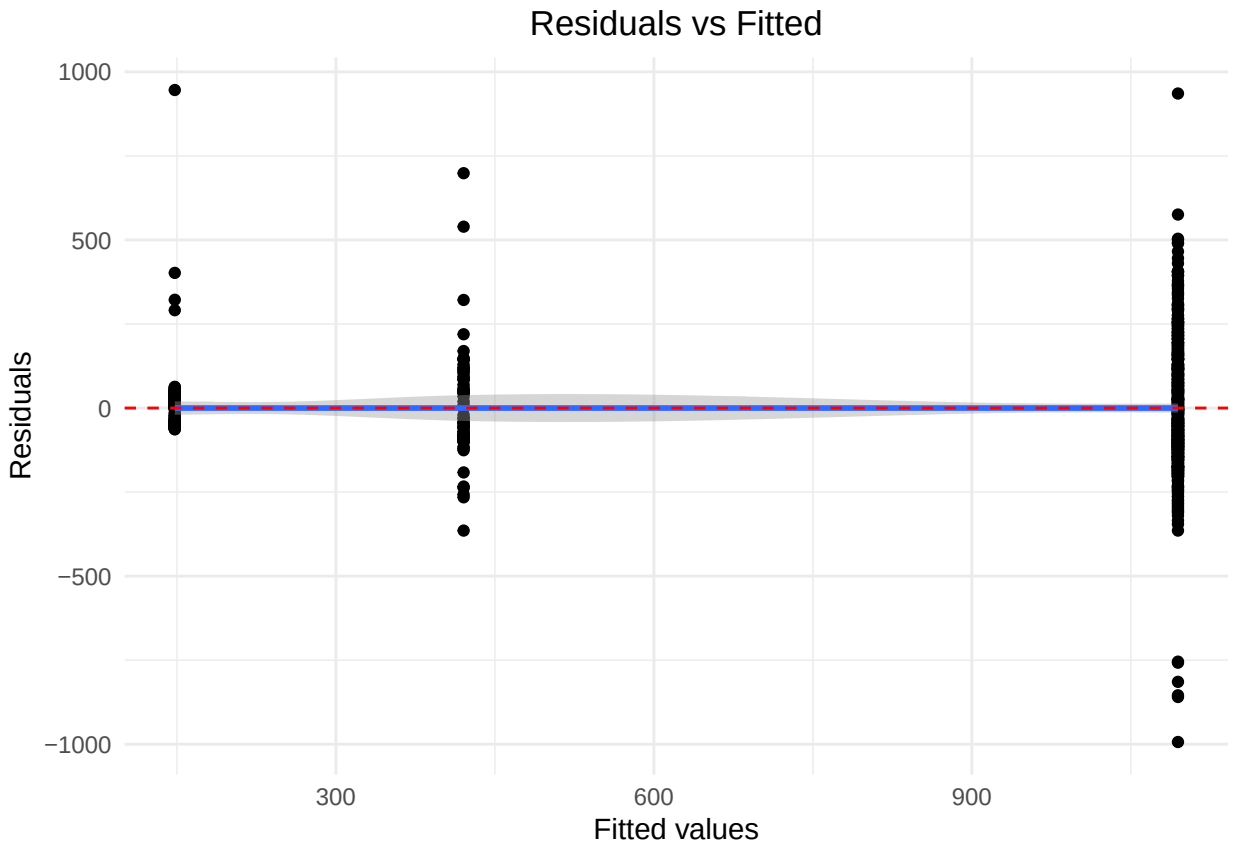
```
# Data
data(Hawks)

# Use the minimal theme
theme_set(theme_minimal())

# Disable warnings (they clutter the document)
options(warn=-1)

# Get rid of rows with NA values in weight. I could probably replace the missing
# values with averages or medians (using imputer, etc), but in this case, will
# just delete these rows.
#
# NOTE: We only need the following columns: Species, Weight
```

```
# We first extract the relevant columns
Hawks = Hawks[, c("Species", "Weight")]

# We now remove the rows where the value of the Weight column is NA
Hawks = Hawks[which(!is.na(Hawks$Weight)), ]

# Dotplot
ggplot(Hawks, aes(x = Species, y = Weight, fill = Species)) +
    geom_dotplot(binaxis = "y", binwidth = 15,
                 dotsize = 1.25, stackdir = "center") +
    ggtitle("Weights by Species Dotplot") +
    theme(plot.title = element_text(hjust = 0.5))
```




```
# Define the linear regression model
lr = lm (Weight ~ Species, data=Hawks)

# Residuals vs Fitted
ggplot(lr, aes(x = .fitted, y = .resid)) +
    geom_point() +
    geom_smooth() +
    geom_hline(yintercept = 0, col = "red", linetype = "dashed") +
    ggtitle("Residuals vs Fitted") +
    theme(plot.title = element_text(hjust = 0.5)) +
    xlab("Fitted values") +
    ylab("Residuals")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
# Calculate standard deviation values
Hawks %>% group_by (Species) %>% summarize (sd.Weight = sd (Weight))
```

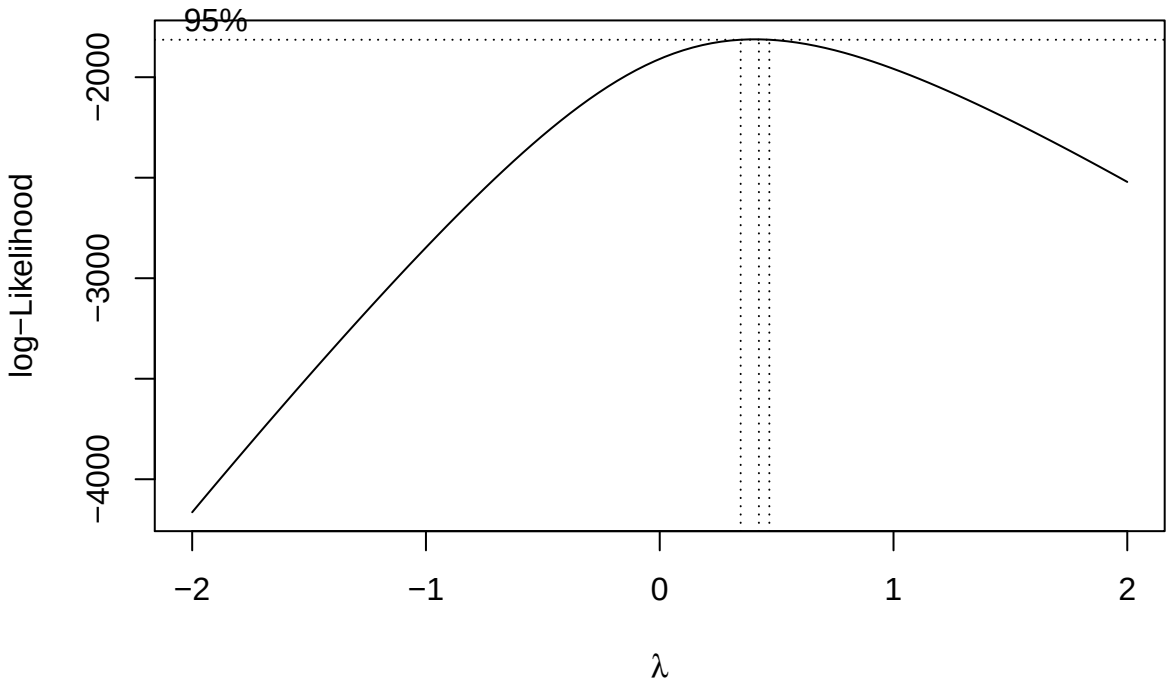
```
## # A tibble: 3 x 2
##   Species sd.Weight
## * <fct>       <dbl>
## 1 CH          162.
## 2 RT          189.
## 3 SS           80.7
```

(a) In terms of weight, red-tailed hawks seem to have the largest weight, followed by Cooper's hawks species, and finally, the sharp-shinned hawks. It seems like the variance is approximately equal across all three groups and hence, equal variance assumption of ANOVA is met. In order to verify that this is true, we have also shown The Residuals vs Fitted plot which confirms constant variance across the groups.

(b) ANOVA assumes that the population standard deviations for all levels are equal. The approximate standard deviation values we have gotten are 162, 189, 80.7 for Cooper's, red-tailed, and sharp-shinned hawks respectively. These values are not approximately equal and hence, it does not meet at least one of assumptions for performing ANOVA.

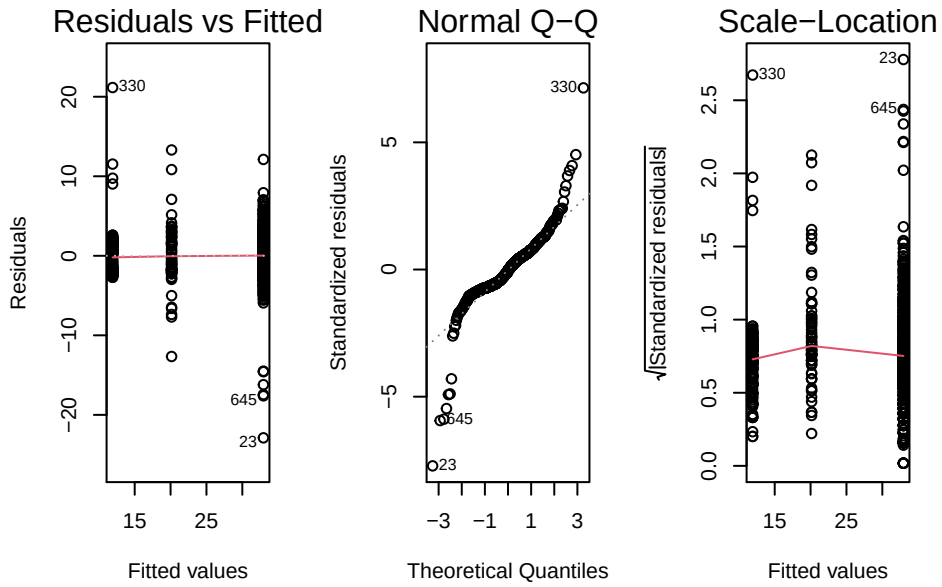
Now, let us find a suitable response transformation and and do a one-way ANOVA on that transformed scale. We will also include residual analysis and all pairwise group comparisons.

```
boxcox(lr)
```



Box-Cox plot has the form of a flipped parabola ($f(x) = -x^2$). The λ (lambda) value is approximately 0.3 and is between 0.25 and 0.45 with 95% confidence. These observations suggest the non-normality of errors in the linear model. Therefore, it is reasonable to perform the square root transformation of the response variable (Weight). This would help us normalize the errors as well as address the non-linearity of the distribution.

```
# Perform the square root transform
lrt = lm (sqrt(Weight) ~ Species, data=Hawks)

# Residual analysis
par (mfrow=c(1,3))
plot (lrt, which=1:3)
```



```
par (mfrow=c(1,1))

# ANOVA
emmeans (lrt, pairwise ~ Species)
```

```
## Note: Use 'contrast(regrid(object), ...)' to obtain contrasts of back-transformed estimates
```

```
## $emmeans
##  Species emmean    SE  df lower.CL upper.CL
##  CH        20.1 0.355 895     19.4     20.8
##  RT        32.9 0.124 895     32.7     33.2
##  SS        11.9 0.185 895     11.6     12.3
##
## Results are given on the sqrt (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE  df t.ratio p.value
##  CH - RT    -12.80 0.376 895 -34.053 <.0001
##  CH - SS      8.22 0.400 895  20.544 <.0001
##  RT - SS     21.02 0.223 895  94.192 <.0001
##
## Note: contrasts are still on the sqrt scale
## P value adjustment: tukey method for comparing a family of 3 estimates
```

After performing the square root transformation, we have already gotten better with all plots. Residuals vs Fitted is better as the black line follows the red dotted line. Normal Q-Q plot has also improved and the variance is nearly constant. The third plot $\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows that the residuals are spread equally along the ranges of predictors and that the variance is nearly constant. Besides, the square root of Standardized Residuals show the decreasing trend.

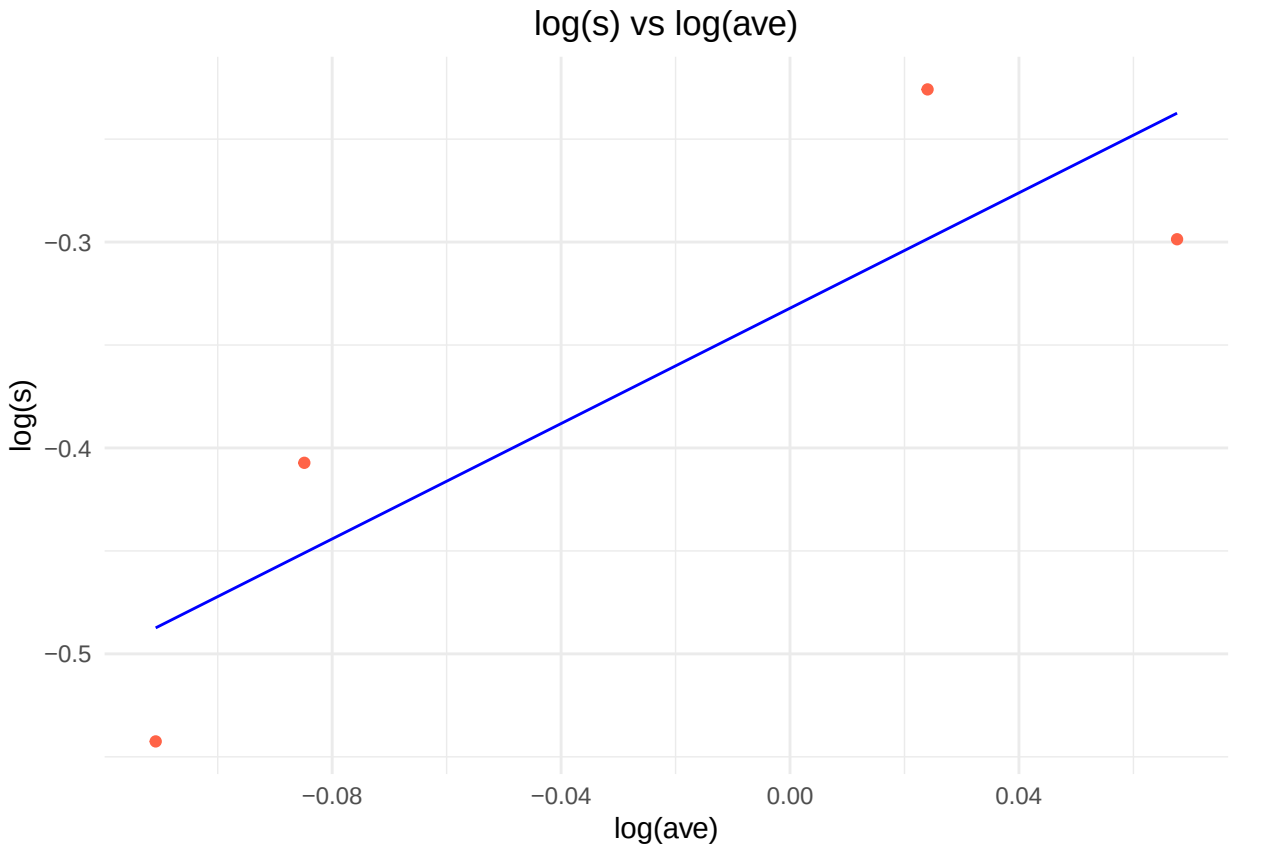
NOW, INTERPRET ANOVA RESULTS

Exercise 5.42

Do Exercise 5.42 as stated. Also do a Box-Cox analysis to see what response transformation it would recommend. Choose a response transformation and do a one-way ANOVA on the transformed response (Carat) vs. color. Analyze the residuals and report all pairwise comparisons.

```
data ("Diamonds2")
Diam.summ = Diamonds2 %>% group_by (Color) %>% summarize (mean.Carat = mean (Carat),
                                                          sd.Carat = sd (Carat))
Diam.summ$log.mean = log10 (Diam.summ$mean.Carat)
Diam.summ$log.sd   = log10 (Diam.summ$sd.Carat)

diamond_lm = lm(Diam.summ$log.sd ~ Diam.summ$log.mean)

ggplot() +
    geom_point(aes(x = Diam.summ$log.mean, y = Diam.summ$log.sd),
               color="tomato") +
    geom_line(aes(x = Diam.summ$log.mean,
                  y = predict(diamond_lm, list(Diam.summ$log.sd))),
              color="blue") +
    ggtitle("log(s) vs log(ave)") +
    theme(plot.title = element_text(hjust = 0.5)) +
```

```
    xlab("log(ave)") +
    ylab("log(s)")
```




```
Diam.summ
```

```
## # A tibble: 4 x 5
##   Color mean.Carat sd.Carat log.mean log.sd
##   <fct>      <dbl>    <dbl>    <dbl>  <dbl>
## 1 D          0.822    0.392  -0.0849 -0.407
## 2 E          0.775    0.287  -0.111  -0.543
## 3 F          1.06     0.594   0.0240 -0.226
## 4 G          1.17     0.503   0.0676 -0.299
```

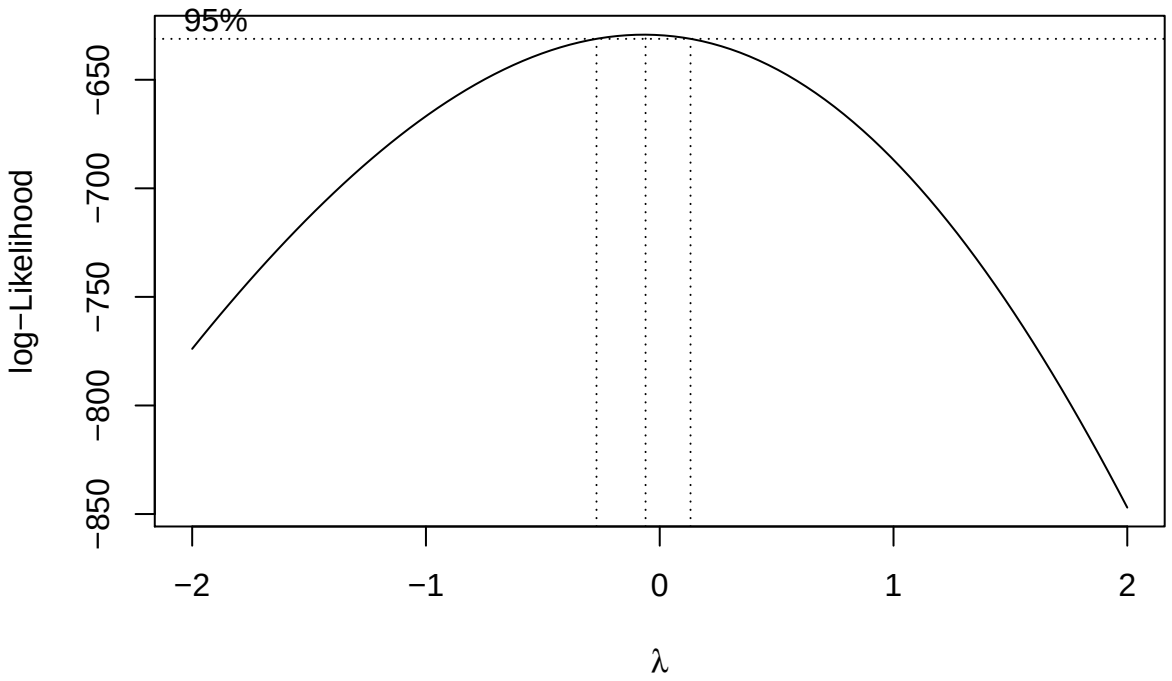
(a) The points seem to suggest a $\frac{1}{x}$ function. However, it is possible to eyeball a line.

(b) We can eyeball the slope. It is in the range 1.4 - 1.5. We can say that it is approximately 1.45. After making a model and estimating the slope, we got that the slope is approximately 1.4008 $\approx$ 1.4 which is not too far from our initial observation.

(c) Since the slope is approximately 1.4, we get $1 - \text{slope} = 1 - 1.4 = -0.4$. Since the value is -0.4, the reciprocal transformation of the form $\frac{1}{\sqrt{\text{response}}}$ is suggested.

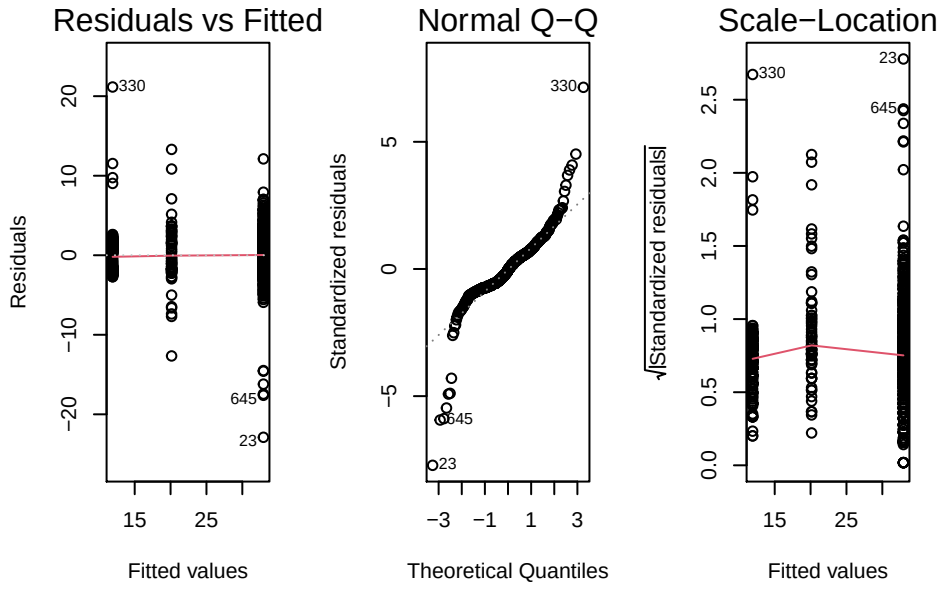
We will now proceed by performing the Box-Cox analysis to see what response transformation it would recommend. We will then choose a response transformation and do a one-way ANOVA on the transformed response (Carat) vs. color. Analyze the residuals and report all pairwise comparisons.

```
carat_lm = lm(Carat ~ Color, Diamonds2)
boxcox(carat_lm)
```



Box-Cox plot has the form of a flipped parabola ($f(x) = -x^2$). The λ (lambda) value is approximately 0 and is between -0.2 and 0.15 with 95% confidence. These observations suggest the non-normality of errors in the linear model. Therefore, it is reasonable to perform the log transformation of the response variable (Carat). This would help us normalize the errors as well as address the non-linearity of the distribution.

```
# Perform the square root transform
lrt = lm (sqrt(Weight) ~ Species, data=Hawks)

# Residual analysis
par (mfrow=c(1,3))
plot (lrt, which=1:3)
```




```
par (mfrow=c(1,1))
```

```r
# ANOVA
emmeans (lrt, pairwise ~ Species)
```

```
## Note: Use 'contrast(regrid(object), ...)' to obtain contrasts of back-transformed estimates
```

```
## $emmeans
##  Species emmean    SE  df lower.CL upper.CL
##  CH        20.1 0.355 895     19.4     20.8
##  RT        32.9 0.124 895     32.7     33.2
##  SS        11.9 0.185 895     11.6     12.3
##
## Results are given on the sqrt (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE  df t.ratio p.value
##  CH - RT    -12.80 0.376 895 -34.053 <.0001
##  CH - SS      8.22 0.400 895  20.544 <.0001
##  RT - SS     21.02 0.223 895  94.192 <.0001
##
## Note: contrasts are still on the sqrt scale
## P value adjustment: tukey method for comparing a family of 3 estimates
```

After performing the square root transformation, we have already gotten better with all plots. Residuals vs Fitted is better as the black line follows the red dotted line. Normal Q-Q plot has also improved and the variance is nearly constant. The third plot ($\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows that the residuals are spread equally along the ranges of predictors and that the variance is nearly constant. Besides, the square root of Standardized Residuals show the decreasing trend.

NOW, INTERPRET ANOVA RESULTS