Chapter 5 Homework Part B

David Oniani

February 22, 2021

Contents

Setup 2
Exercise 5.32 2
Exercise 5.42 7

Setup

```
library(MASS)
library(Stat2Data)
library(dplyr)
```

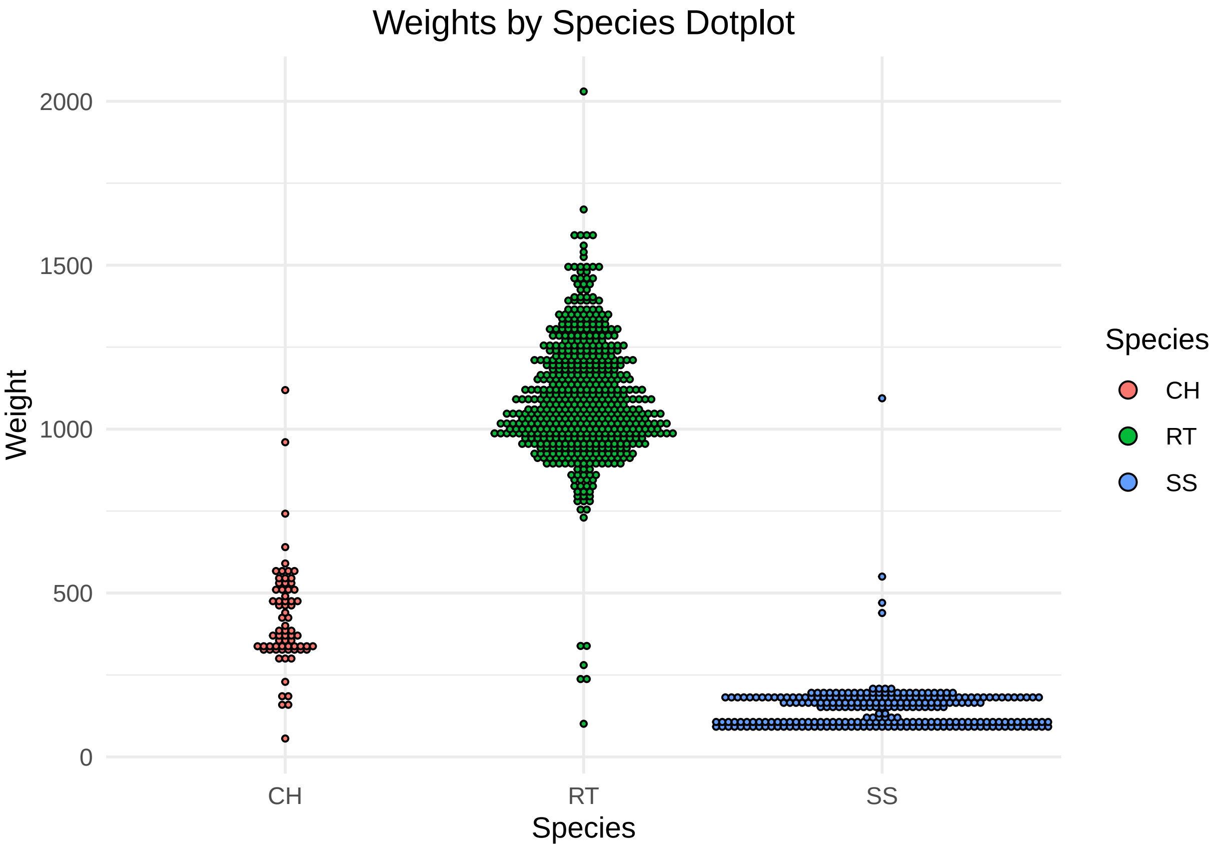
```
##
## Attaching package: 'dplyr'

## 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)

# Use the minimal theme
theme_set(theme_minimal())

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

Exercise 5.32

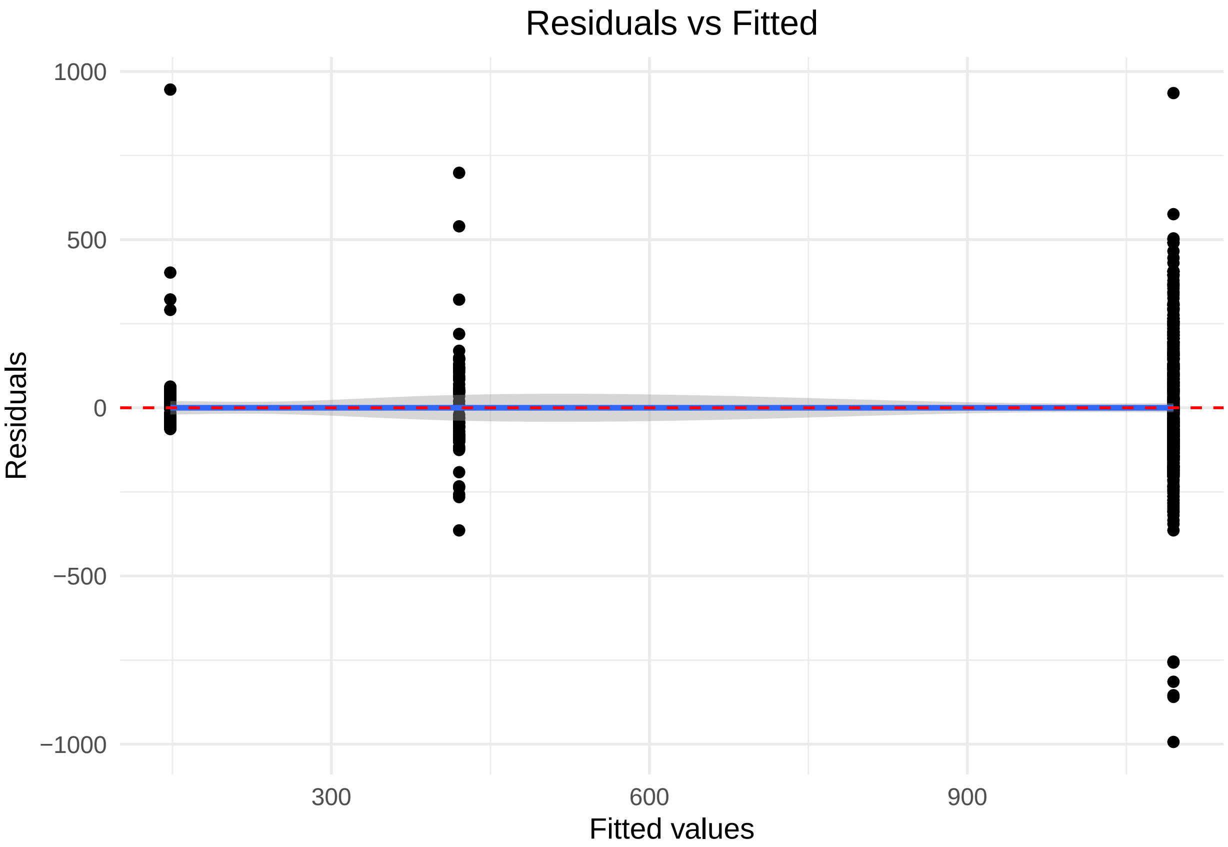
Do Exercise 5.32 as stated. Also, find a suitable response transformation and and do a one-way ANOVA on that transformed scale. Include residual analysis and all pairwise group comparisons.

```
# Data
data(Hawks)

# 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'
```



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

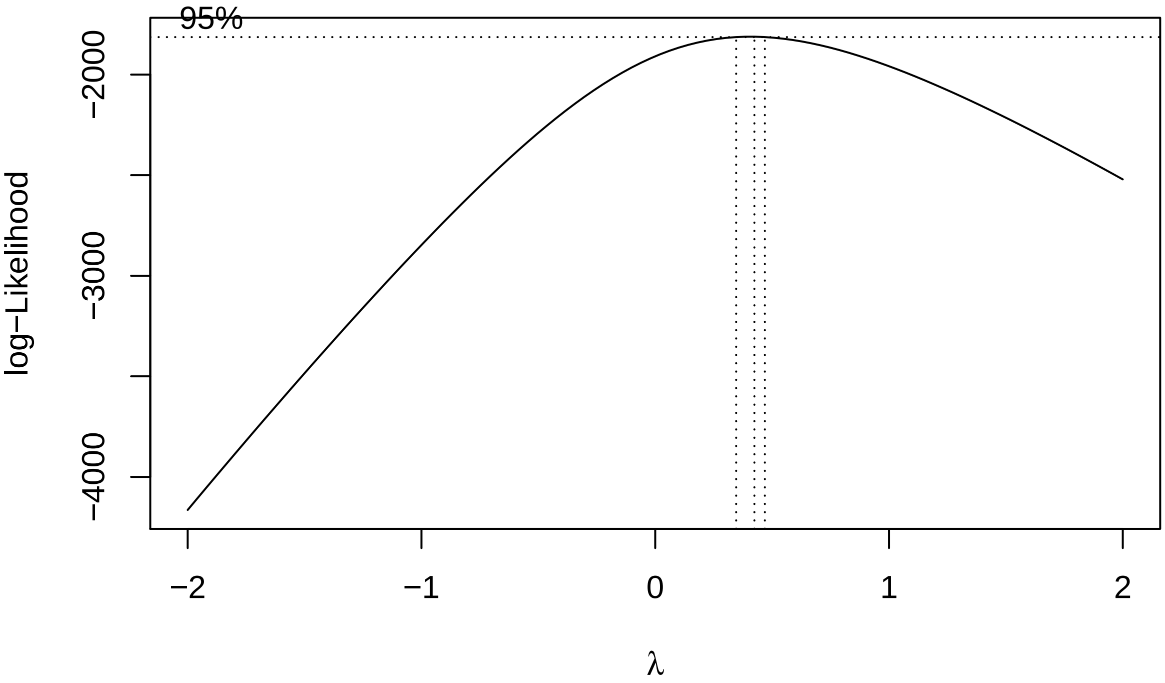
```
## # A tibble: 3 x 2
##   Species sd.Weight
## * <fct>       <dbl>
## 1 CH          162.
## 2 RT          189.
## 3 SS           80.7
```

(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 (the red line approximately follows the dotted line).

(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)
```



```
summary(lr)
```

```
##
## Call:
## lm(formula = Weight ~ Species, data = Hawks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -993.43  -80.49  -14.93   55.57  946.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   420.49      19.54   21.52   <2e-16 ***
## SpeciesRT     673.94      20.70   32.56   <2e-16 ***
## SpeciesSS    -272.52      22.05  -12.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 163.5 on 895 degrees of freedom
## Multiple R-squared:  0.8753, Adjusted R-squared:  0.875
## F-statistic:  3140 on 2 and 895 DF,  p-value: < 2.2e-16
```

Box-Cox plot has the form of a flipped parabola ($f(x) = -x^2$). The λ (lambda) value is approximately 0.4 and is between 0.45 and 0.5 with 95% confidence (default is 95%). 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.

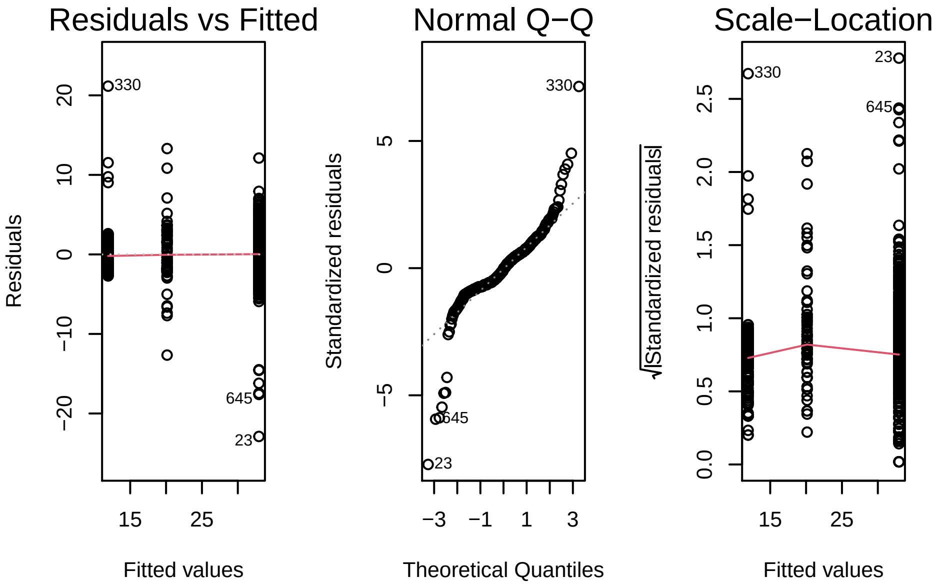
It is also important to note that approximately 87.5% of the variation in Weight is explained by Species.

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

```
summary(lr)
```

```
##
## Call:
## lm(formula = sqrt(Weight) ~ Species, data = Hawks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.8904  -1.7955  -0.1182   1.6285  21.1541
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.1443     0.3547   56.79   <2e-16 ***
## SpeciesRT    12.7960     0.3758   34.05   <2e-16 ***
## SpeciesSS    -8.2227     0.4003  -20.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.968 on 895 degrees of freedom
## Multiple R-squared:  0.911,  Adjusted R-squared:  0.9108
## F-statistic:  4581 on 2 and 895 DF,  p-value: < 2.2e-16
```

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



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

# ANOVA
s = summary(emmeans(lr, pairwise ~ Species), infer = c(T, T))
```

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

```
s
```

```
## $emmeans
##  Species emmean    SE  df lower.CL upper.CL t.ratio p.value
##  CH        20.1 0.355 895     19.4     20.8  56.794 <.0001
##  RT        32.9 0.124 895     32.7     33.2 265.478 <.0001
##  SS        11.9 0.185 895     11.6     12.3  64.277 <.0001
##
## Results are given on the sqrt (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE  df lower.CL upper.CL t.ratio p.value
##  CH - RT    -12.80 0.376 895   -13.68   -11.91 -34.053 <.0001
##  CH - SS      8.22 0.400 895     7.28     9.16  20.544 <.0001
##  RT - SS     21.02 0.223 895    20.49    21.54  94.192 <.0001
##
## Note: contrasts are still on the sqrt scale
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
## 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 and the variance is nearly constant. Normal Q-Q plot has also improved and the distribution is more normal. The Scale-Location plot shows that the residuals are spread equally along the ranges of predictors and that the variance is nearly constant. Besides, the absolute value of the square root of Standardized Residuals show the decreasing trend.

We can also see the adjusted R-squared value of 0.9108 which means that approximately 91.08% of variation in $\sqrt{Weight}$ is explained by Species which is a good improvement over the previous 87.5%.

Our null hypothesis is that the differences of means across the groups are zero and the alternative hypothesis is that at least one of differences of means is different from 0. The $contrasts table shows the following results:

```
s$contrast
```

```
##  contrast estimate    SE  df lower.CL upper.CL t.ratio p.value
##  CH - RT    -12.80 0.376 895   -13.68   -11.91 -34.053 <.0001
##  CH - SS      8.22 0.400 895     7.28     9.16  20.544 <.0001
##  RT - SS     21.02 0.223 895    20.49    21.54  94.192 <.0001
##
## Note: contrasts are still on the sqrt scale
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
## P value adjustment: tukey method for comparing a family of 3 estimates
```

We get that the difference of means between Cooper's hawks and red-tailed hawks is -12.80 (could have phrased it as "difference between the mean of red-tailed hawks and Cooper's hawks is 12.80") which is statistically significant with the p-value less than 0.0001. The difference between the means of Cooper's and sharp-shinned hawks is 8.22 which is also statistically significant with the p-value < 0.0001. The difference of means between red-tailed and sharp-shinned hawks is 21.02 with the p-value < 0.0001. Hence, we reject the null and accept the alternative. Finally, we conclude that the variance between the three groups is not the same.

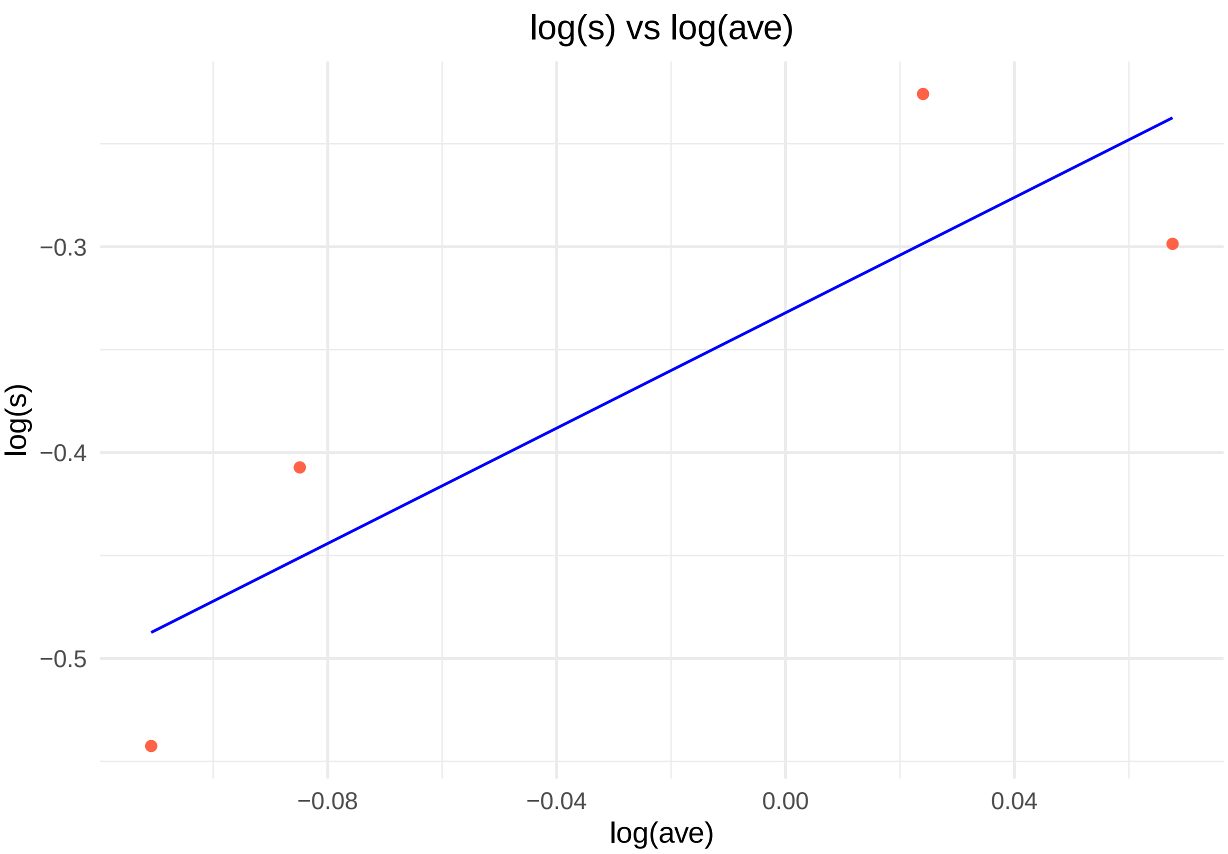
Exercise 5.42

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)

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
```

```
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)")
```



(a) The points seem to suggest a $\log x$ 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 square root transformation of the form $\frac{1}{\sqrt{\text{response}}}$ is suggested.

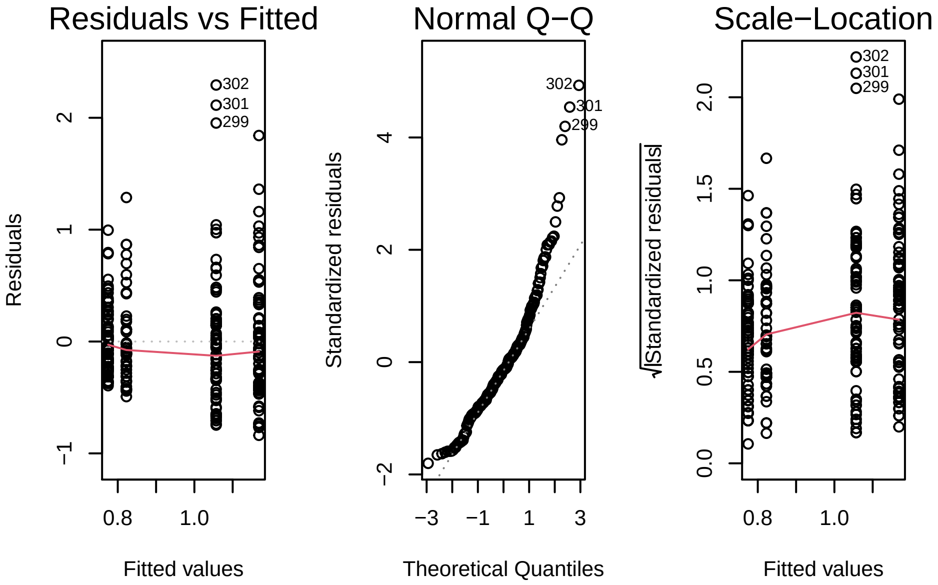
Let us now perform the suggested transformation.

```r
diamond_lm_transform = lm(1 / sqrt(Carat) ~ Color, Diamonds2)
confint(diamond_lm_transform)
```

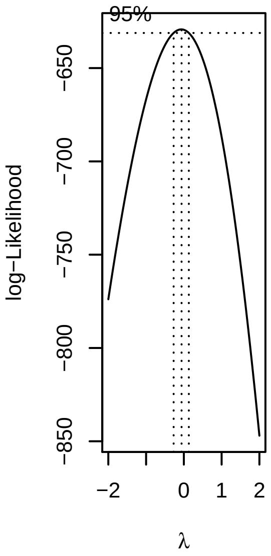
```
##                   2.5 %      97.5 %
## (Intercept)  1.11728401  1.25149329
## ColorE      -0.08233953  0.08922519
## ColorF      -0.18802192 -0.01838112
## ColorG      -0.27745738 -0.10744802
```

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. Finally, we will analyze the residuals and report all pairwise comparisons.

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

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



```
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.

It is important to note that the Residuals vs Fitted plot does not look good (red line does not exactly follow the dotted line). Additionally, Normal Q-Q also has some problems toward the tail of the plot.

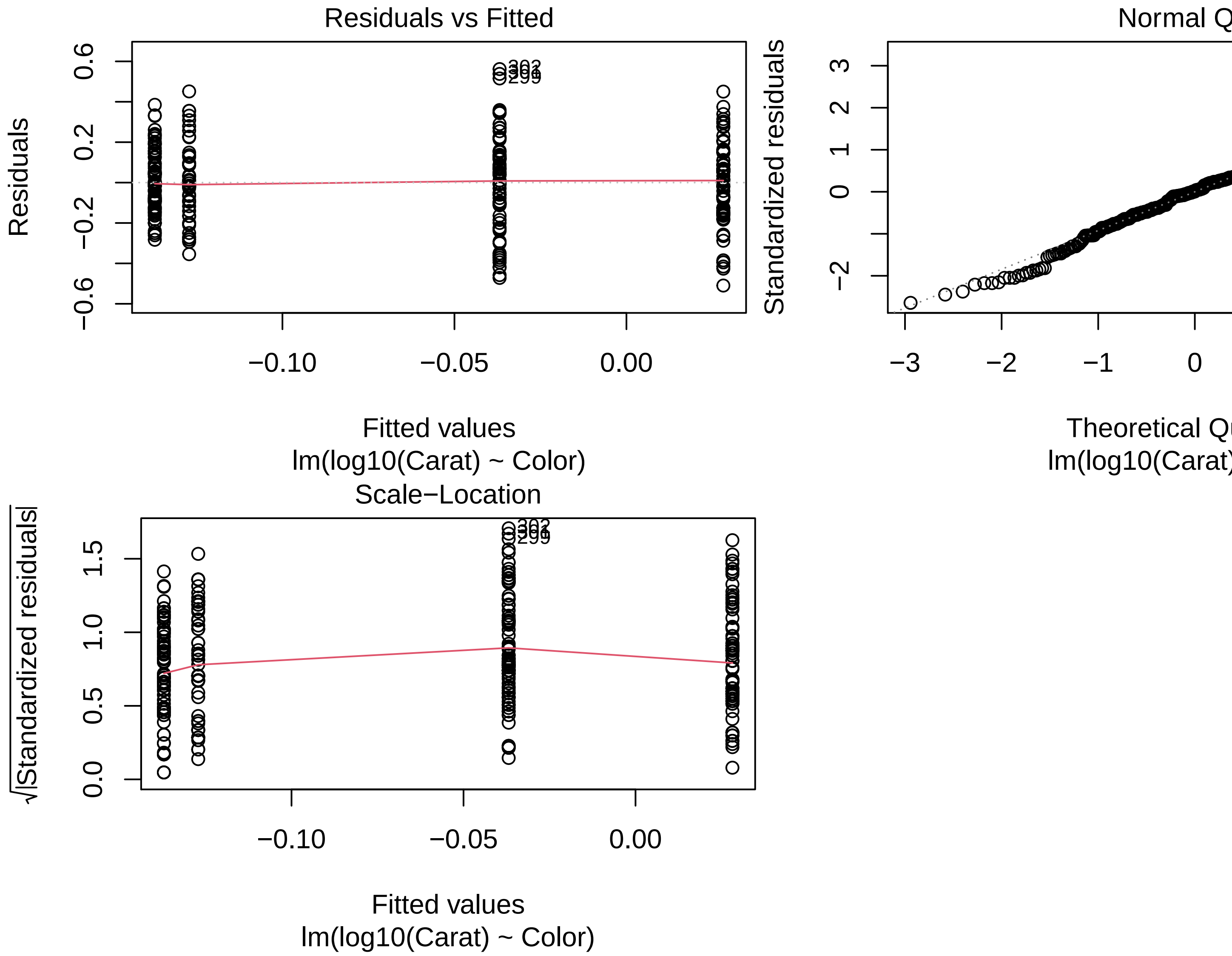
```
# Perform the square root transform
lr = lm(log10(Carat) ~ Color, Diamonds2)

lr$cont
```

```
## $Color
## [1] "contr.treatment"
```

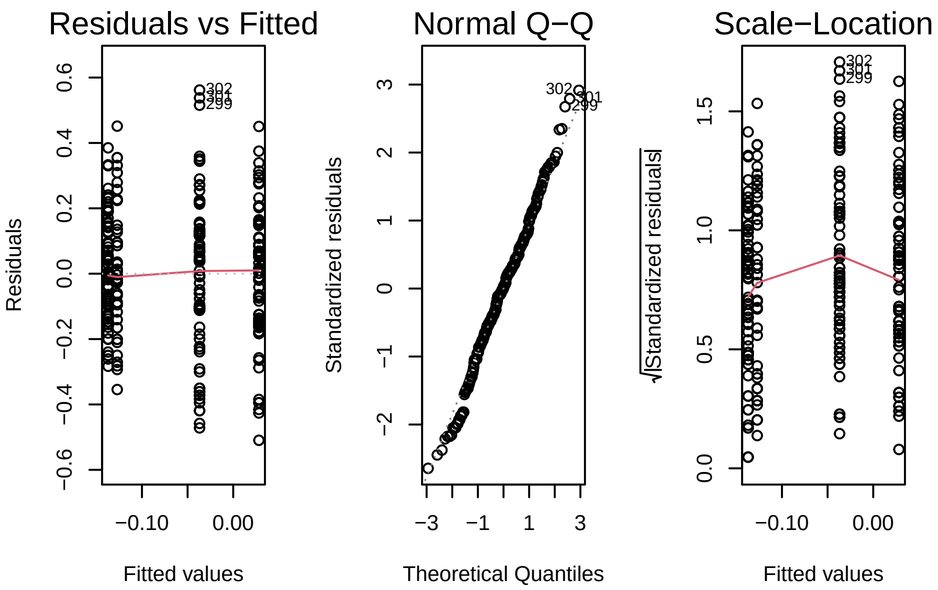
```
confint(lr)
```

```
##                   2.5 %      97.5 %
## (Intercept) -0.18002878 -0.07421724
## ColorE      -0.07761417  0.05764869
## ColorF       0.02338046  0.15712647
## ColorG       0.08827509  0.22231168
```

```
# Residual analysis
plot(lr, which = 1:3)
```



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



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

# ANOVA
s = summary(emmeans(lr, pairwise ~ Color), infer = c(T, T))
```

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

```
s
```

```
## $emmeans
##  Color  emmean     SE  df lower.CL upper.CL t.ratio p.value
##  D     -0.1271 0.0269 303  -0.1800 -0.07422 -4.728  <.0001
##  E     -0.1371 0.0214 303  -0.1792 -0.09498 -6.404  <.0001
##  F     -0.0369 0.0208 303  -0.0778  0.00403 -1.774  0.0771
##  G      0.0282 0.0209 303  -0.0130  0.06931  1.347  0.1788
##
## Results are given on the log10 (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate     SE  df lower.CL upper.CL t.ratio p.value
##  D - E     0.00998 0.0344 303  -0.0788  0.09877  0.290  0.9915
##  D - F    -0.09025 0.0340 303  -0.1780 -0.00246 -2.656  0.0413
##  D - G    -0.15529 0.0341 303  -0.2433 -0.06731 -4.560  <.0001
##  E - F    -0.10024 0.0298 303  -0.1773 -0.02315 -3.359  0.0049
##  E - G    -0.16528 0.0299 303  -0.2426 -0.08797 -5.523  <.0001
##  F - G    -0.06504 0.0295 303  -0.1412  0.01112 -2.206  0.1238
##
## Note: contrasts are still on the log10 scale
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 4 estimates
## P value adjustment: tukey method for comparing a family of 4 estimates
```

After performing the log transformation, we have already gotten better with all plots. Residuals vs Fitted is better as the red line follows the black dotted line. Normal Q-Q plot has also improved and the normality assumption for ANOVA is clearly met. The Scale-Location plot shows that the residuals are spread equally along the ranges of predictors and that the variance is nearly constant. Furthermore, the square root of

Standardized Residuals show the decreasing trend.

Our null hypothesis is that the differences of means across the groups are zero and the alternative hypothesis is that at least one of differences of means is different from 0. The $contrasts table shows the following results:

```
s$contrast
```

```
##  contrast estimate     SE  df lower.CL upper.CL t.ratio p.value
##  D - E     0.00998 0.0344 303  -0.0788  0.09877  0.290  0.9915
##  D - F    -0.09025 0.0340 303  -0.1780 -0.00246 -2.656  0.0413
##  D - G    -0.15529 0.0341 303  -0.2433 -0.06731 -4.560  <.0001
##  E - F    -0.10024 0.0298 303  -0.1773 -0.02315 -3.359  0.0049
##  E - G    -0.16528 0.0299 303  -0.2426 -0.08797 -5.523  <.0001
##  F - G    -0.06504 0.0295 303  -0.1412  0.01112 -2.206  0.1238
##
## Note: contrasts are still on the log10 scale
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 4 estimates
## P value adjustment: tukey method for comparing a family of 4 estimates
```

Only two of the differences between means yields statistically significant results and these are differences D - G (with -0.15529) and E - G (0.16528) where p-value < 0.0001 in both cases. Also, in D - F and E - F we have the p-values of 0.0413 and 0.0049 respectively. Therefore, all of them are statistically significant. Hence, we reject the null and accept the alternative hypothesis suggesting that means across the groups are different.