ENV 710: Lab 7

Jiahuan Li

Spring 2023

Problem 1

Hypothesis

The null hypothesis is that the average weight of confiscated elephant tusks is the same across the years 1970, 1990, and 2010. While the alternative hypothesis is that the average weight of confiscated elephant tusks has decreased over time in the three year groups.

Model

To test this hypothesis, we will use a one-way ANOVA design with year as the factor. That is because there is only one independent variable in the present model. Besides, the limited number of subjects in this study precludes a random effect analysis. Consequently, the model is constructed as follows:

```
##
## Call:
## lm(formula = Tusk.kg ~ factor(Year), data = Tusk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6964 -0.9710 -0.0348  0.9109  3.9427
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        11.7150     0.3438   34.08   <2e-16 ***
## factor(Year)1990   -8.7916     0.4862  -18.08   <2e-16 ***
## factor(Year)2010   -9.5400     0.4862  -19.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.537 on 57 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.8894
## F-statistic: 238.1 on 2 and 57 DF,  p-value: < 2.2e-16
```

The ANOVA results show that there is a statistically significant difference in the mean weight of elephant tusks across the three years ($R^2 = 0.889$, $F_{2,57} = 238.1$, $p <$ 2.2e-16). The R-squared value indicates that the model explains a high percentage of the variance in the response variable. The F-statistic with a p-value less than 0.05 indicates that the model is statistically significant.

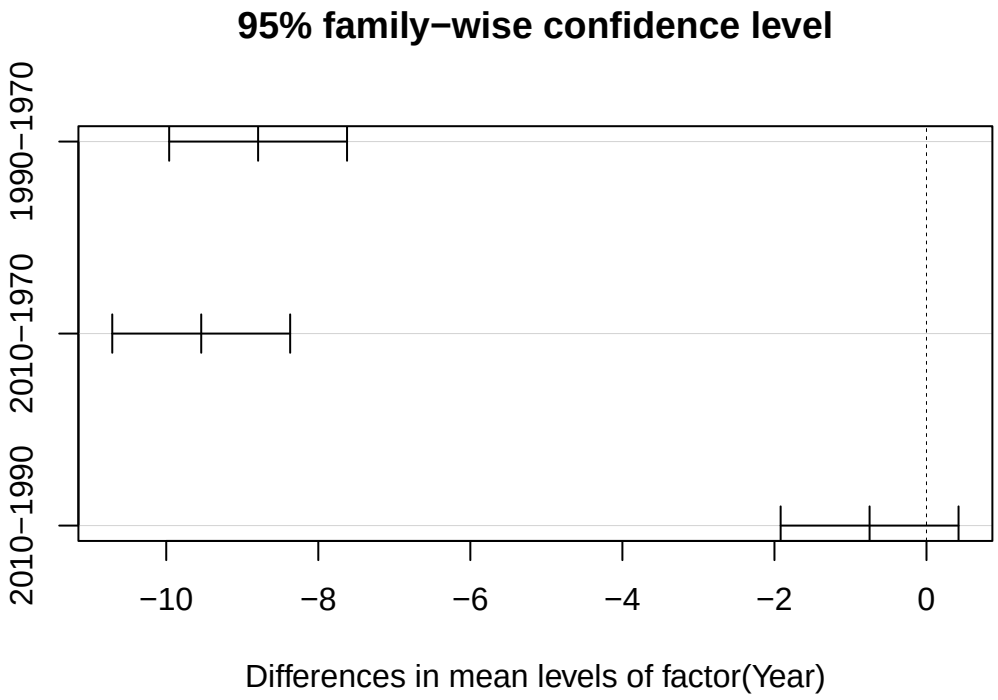
Based on the analysis of the data, it can be inferred that the hypothesis positing a decline in the mean weight of the confiscated elephant tusks over time is supported. With the reference average weight of 1970 serves as the intercept, the output reveals that the slopes for both 1990 and 2010 are statistically significant (p-value $<$ 0.05) with negative values. Furthermore, the slope for 2010 (-9.54) is more negative than that of 1990 (-8.79), indicating a more significant decrease in the average weight of the tusks over time. Therefore, the findings

suggest that the average weight of confiscated elephant tusks has significantly decreased over the three year groups, supporting the alternative hypothesis.

Post-hoc test

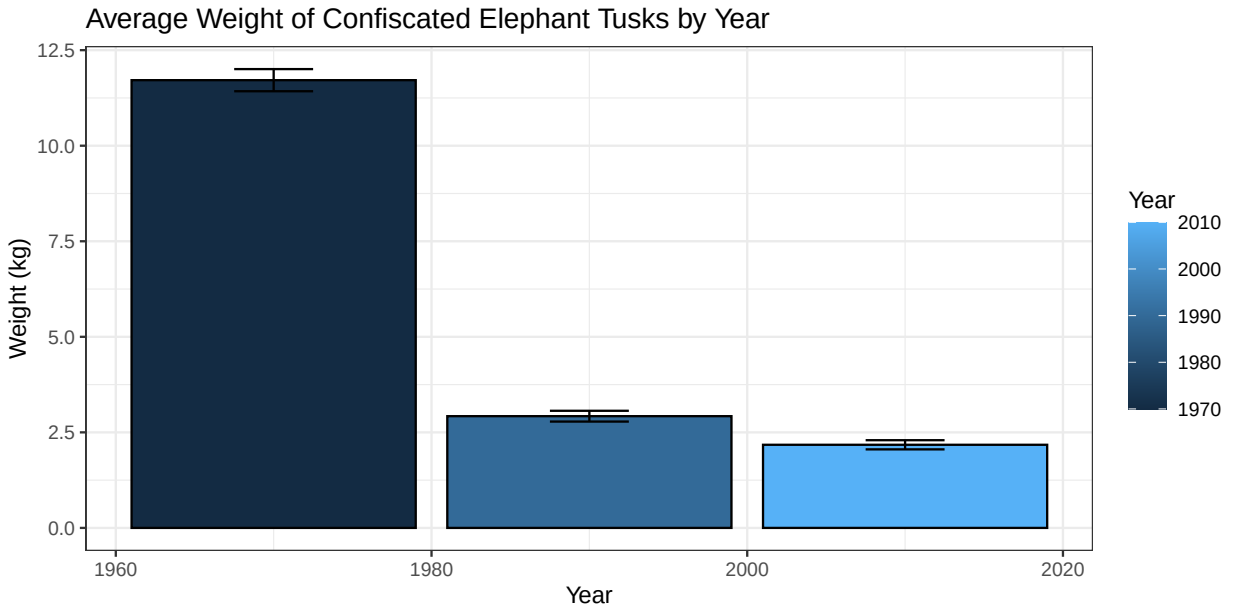
To follow up on this result, I have performed a Tukey's HSD test to determine which pairs of years have significantly different mean weights.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $`factor(Year)`
##                 diff        lwr        upr     p adj
## 1990-1970 -8.7915546  -9.961482 -7.6216274 0.0000000
## 2010-1970 -9.5400032 -10.709930 -8.3700759 0.0000000
## 2010-1990 -0.7484485  -1.918376  0.4214787 0.2803896
```



In this graph, there is not a statistically significant difference between the yeara 2010 and 1990 as illustrated by the fact that the CI overlaps 0 and the p-value is greater than 0.05. By contrast, there appears to be significant differences in biomass between the years 2010 and 1970 and years 1990 and 1970.

Besides, I have also visualized the results via a barplot of the means of tusk weight and standard errors per year group with error bars.

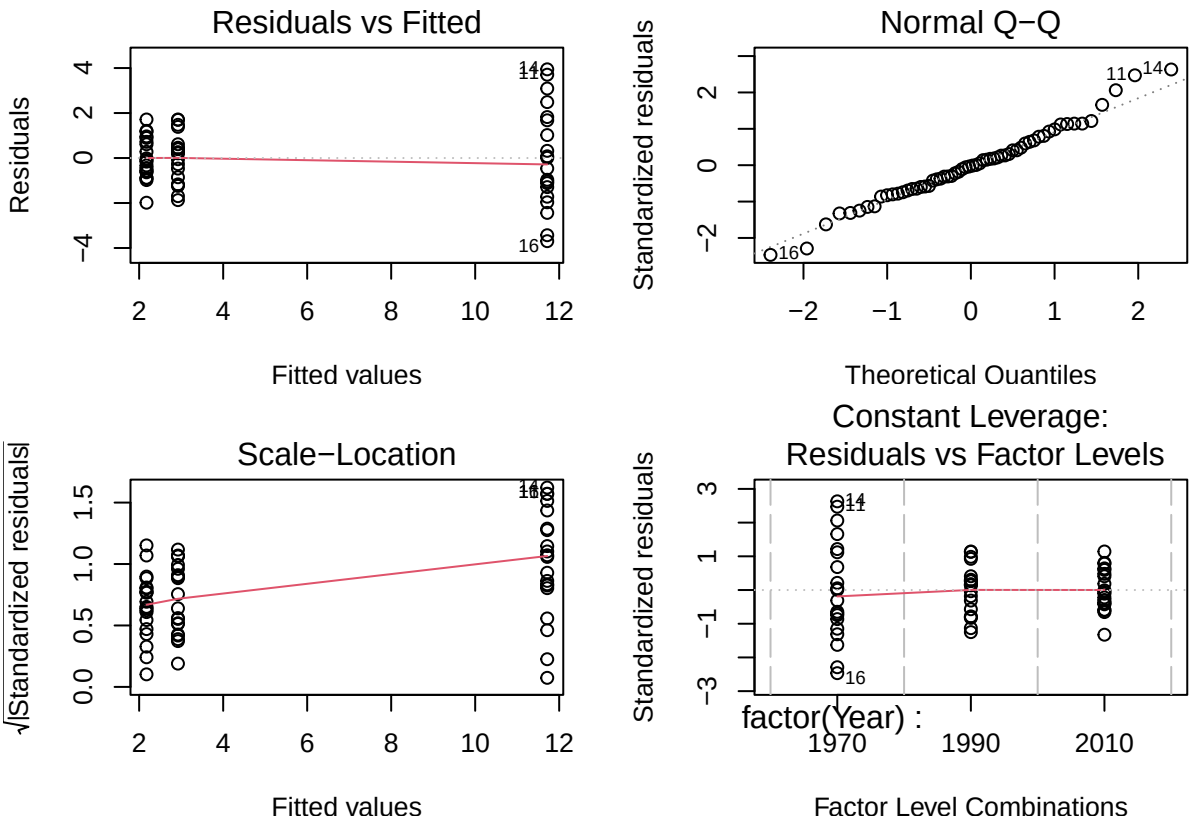


Assumptions

The diagnostic plots were utilized to assess the assumptions of the statistical test. The residual plot showed that the residuals were randomly scattered around the centerline, indicating that they were approximately normally and independently distributed with a mean of 0 and a constant variance. And the QQ plot displayed a linear trend, indicating good normality of the residuals.

Third, however, the figure of the standardized residual plot is worrying since the variance seems to increase with the treatment means, which might be evidence of heteroscedasticity. Further investigation revealed that some of the ratios of the sample standard deviations exceeded the limit of 2. To address this issue, the dependent variable was transformed into a log format and the model was re-run. However, the ratio still did not meet the required criteria. Therefore, I decided to ignore this problem since the other diagnostic plots were satisfactory.

The leverage plot did not show any outliers, although observations 14 and 16 had more leverage than the other observations. Overall, the diagnostic plots suggested that the assumptions of the statistical test were generally met, despite the presence of potential heteroscedasticity.




```
## [1] "The ratios of sample dependent variables' sds are 2.05120566880595 , 1.1793366316645 , 2.419061
```

Problem 2

Hypothesis

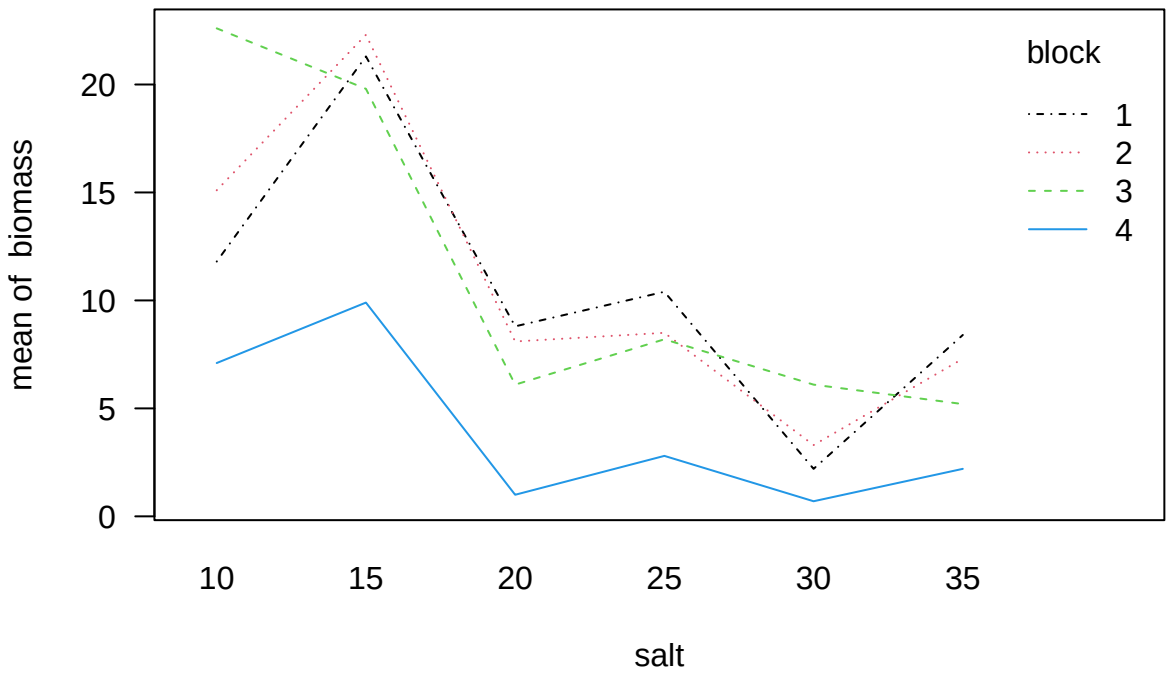
The null hypothesis is that the mean biomass growth is the same across all six levels of salt addition. While the alternative hypothesis is that the mean biomass growth is different across at least one pair of salt addition levels.

Model

In order to test the hypothesis, a complete block design with salt addition as the factor and block as the blocking variable was utilized. It is "complete" because the treatments were assigned separately within each block, ensuring that each treatment was included at least once in every block.

The main reason for not considering the interaction is that the blocks are not the focus of the experiment. The factor of interest is salt addition, and the blocks are simply a part of the experimental design. Moreover, I have also constructed an interaction plot to determine whether the interaction term should be included. However, the lines in the plot are nearly parallel, suggesting that there is no need for that interaction term.

Besides, the block effect is better to be treated as a random effect. However, it is not feasible due to the small number of subjects (less than 6).



Thus, I construct the model as follows:

```
##
## Call:
## lm(formula = biomass ~ factor(salt) + factor(block), data = salt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7000 -1.5729  0.0375  1.4812  6.2500
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     15.5000     1.6981   9.128 1.64e-07 ***
## factor(salt)15   4.1750     1.9608   2.129 0.050205 .
## factor(salt)20  -8.1500     1.9608  -4.157 0.000844 ***
## factor(salt)25  -6.6750     1.9608  -3.404 0.003923 **
## factor(salt)30 -11.0750     1.9608  -5.648 4.63e-05 ***
## factor(salt)35  -8.3750     1.9608  -4.271 0.000669 ***
## factor(block)2   0.2833     1.6009   0.177 0.861893
## factor(block)3   0.8500     1.6009   0.531 0.603237
## factor(block)4  -6.5333     1.6009  -4.081 0.000984 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.773 on 15 degrees of freedom
## Multiple R-squared:  0.8862, Adjusted R-squared:  0.8255
## F-statistic:  14.6 on 8 and 15 DF,  p-value: 8.593e-06
```

Overall, the linear model analysis shows that the model is significant as a whole ($R^2 = 0.826$, $F_{8,15} = 14.6$, $p = 8.593\text{e-}06 < 0.001$). The model's multiple R-squared is 0.8862, indicating that it explains a large proportion of the variation in plant biomass growth. The adjusted R-squared is 0.8255, indicating it has a good fit and is not overfitting the data.

Generally, the analysis shows that both the salt and block factors have a significant effect on plant biomass growth. The estimate for the intercept is 15.5, representing the expected mean plant biomass growth when the salt level is 10 g m-2 and the plot is in block 1.

The estimated coefficients for the factor "salt" represent the difference in mean plant biomass growth between the salt levels and the reference level (10 g m-2). The estimate for salt level 15 is 4.175, indicating that the mean biomass growth is significantly higher with an increase of 4.175 units for this level compared to the reference level (p = 0.05). On the contrary, the estimates for salt levels 20, 25, 30, and 35 are all negative, indicating a decrease in mean biomass growth compared to the reference level. All of these estimates are significant with p-values less than 0.01.

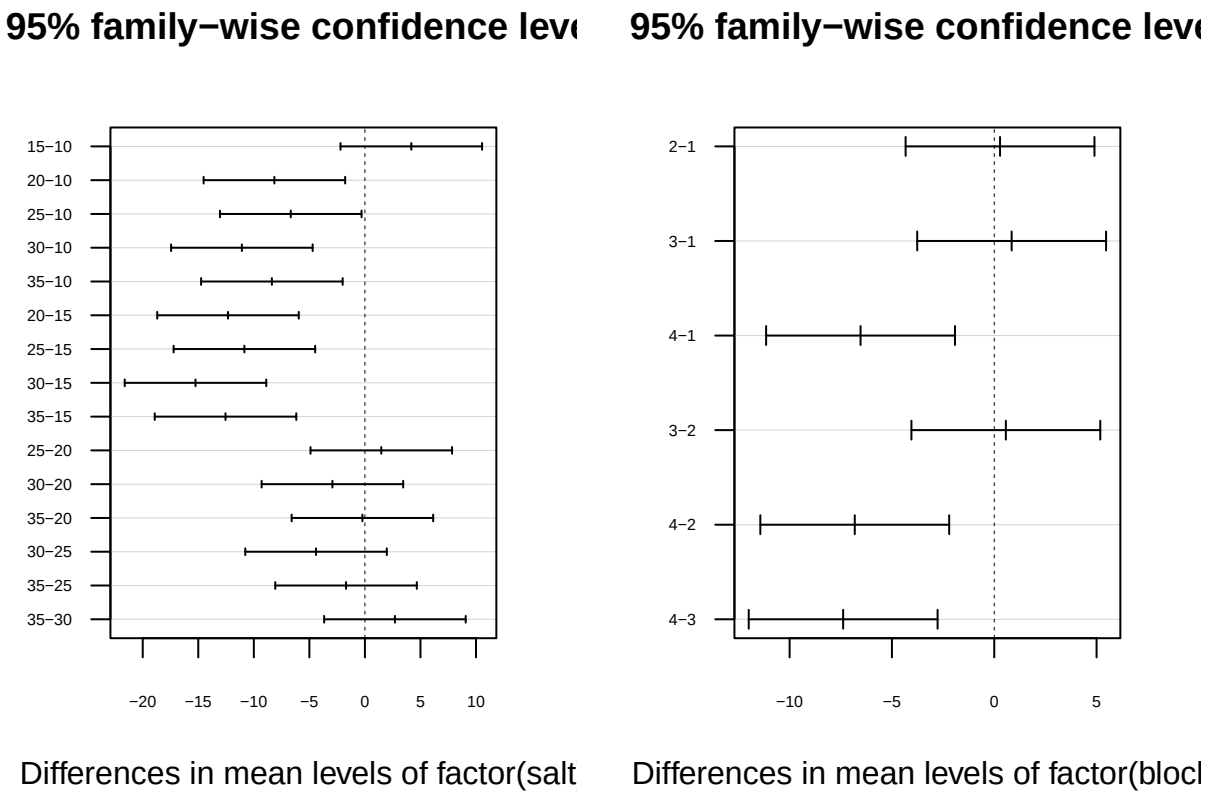
The estimated coefficients for the factor "block" represent the difference in mean plant biomass growth between each block and the reference block (block 1). The estimate for blocks 2 and 3 show that their mean biomass growth is not significantly different from the reference block with p-value of 0.862 and 0.603, respectively. The estimate for block 4 is -6.5333 with a significant p-value, indicating that the mean biomass growth is significantly lower than the reference block by 6.5333 units.

Post-hoc test

Besides the statistically significant main effects of salt and block reported by ANOVA results, I have also studied the group differences using the TukeyHSD test.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $`factor(salt)`
##          diff        lwr        upr     p adj
## 15-10   4.175  -2.195427 10.5454267 0.3242073
## 20-10  -8.150 -14.520427 -1.7795733 0.0089392
## 25-10  -6.675 -13.045427 -0.3045733 0.0374481
## 30-10 -11.075 -17.445427 -4.7045733 0.0005436
## 35-10  -8.375 -14.745427 -2.0045733 0.0071740
## 20-15 -12.325 -18.695427 -5.9545733 0.0001755
## 25-15 -10.850 -17.220427 -4.4795733 0.0006697
## 30-15 -15.250 -21.620427 -8.8795733 0.0000151
## 35-15 -12.550 -18.920427 -6.1795733 0.0001439
## 25-20   1.475  -4.895427  7.8454267 0.9715211
## 30-20  -2.925  -9.295427  3.4454267 0.6742374
## 35-20  -0.225  -6.595427  6.1454267 0.9999965
## 30-25  -4.400 -10.770427  1.9704267 0.2749136
## 35-25  -1.700  -8.070427  4.6704267 0.9487968
## 35-30   2.700  -3.670427  9.0704267 0.7393906
##
## $`factor(block)`
##           diff        lwr       upr     p adj
## 2-1  0.2833333  -4.330839  4.897506 0.9979405
## 3-1  0.8500000  -3.764172  5.464172 0.9501884
## 4-1 -6.5333333 -11.147506 -1.919161 0.0048516
## 3-2  0.5666667  -4.047506  5.180839 0.9841982
```

```
## 4-2 -6.8166667 -11.430839 -2.202494 0.0034273
## 4-3 -7.3833333 -11.997506 -2.769161 0.0017187
```

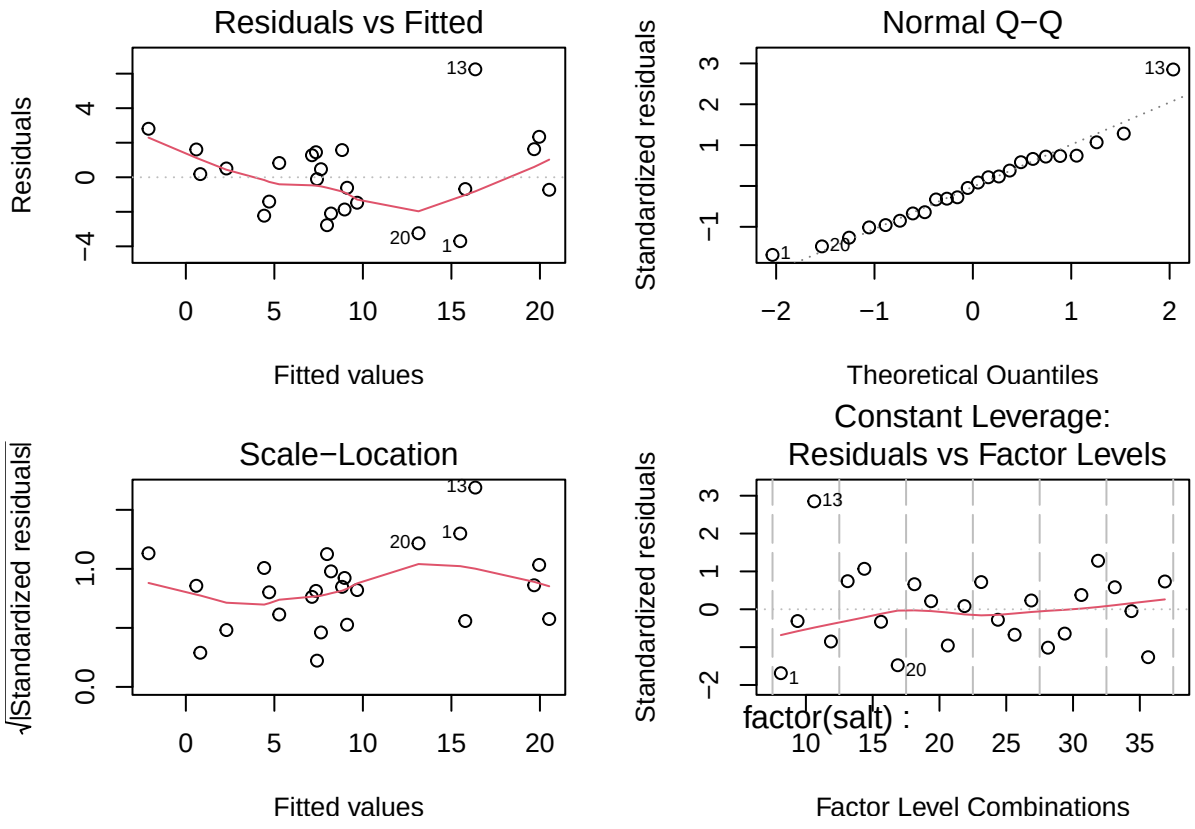


The Tukey tests reflect that there are significant differences between some of the salt addition levels. In other words, the plant biomass growth is significantly different between group pairs (10-20, 25, 30, 35) and pairs (15-20, 25, 30, 35). Thus, I find that the salt level groups of 10 and 15 are not different, and the other four levels 20, 25, 30, and 35 are not different. And from the illustration, it can be easily detected that the block 4 is significantly different than the other three blocks.

Assumptions

At last, I also use the diagnosis plots to evaluate how well the model fits the data. The qq plot suggests that the residuals are normally distributed, which is an indication that the assumption of normality is met. However, the residual plots show a slightly twisted line, which could be a cause for concern. Despite this, the distribution of residuals appears to be random in the first and third graphs.

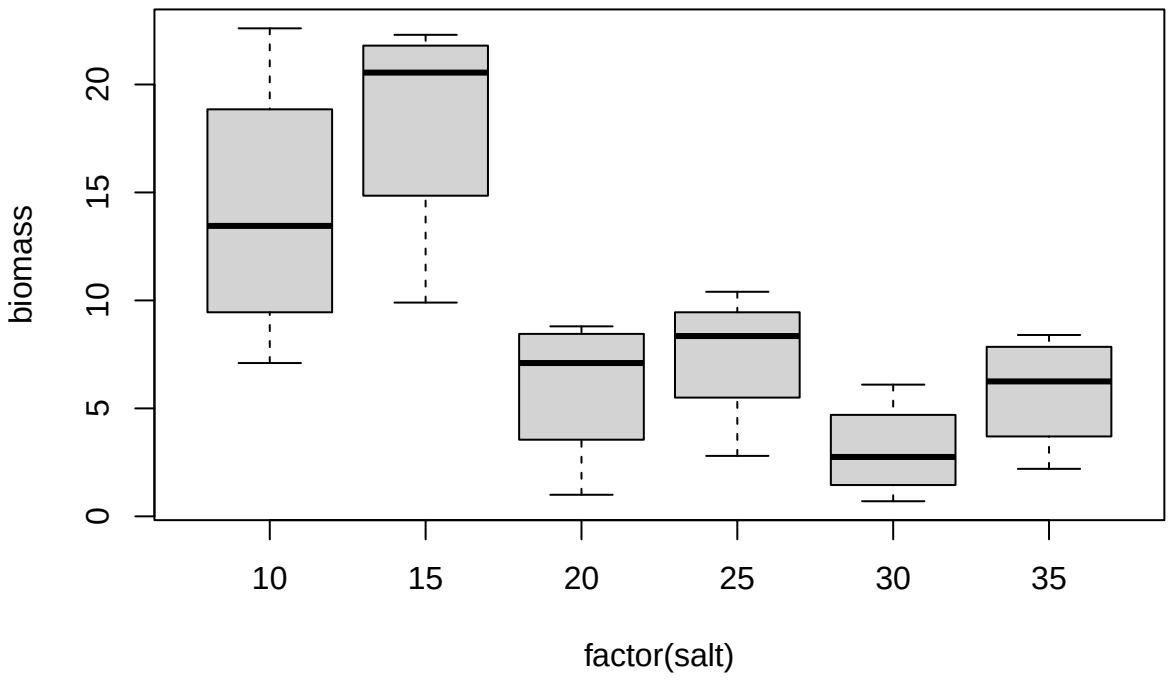
In an effort to address this issue, a log-transformed model was attempted. However, upon examination of the diagnostic plots, it was found that this model does not produce significantly better results. As a result, I decide that the original model setting should be maintained.



Besides, I also use some numeric tests to test whether the distribution of dependent variables meets the model assumptions rather than the residuals which have been tested in the diagnosis plots above. We can find that the p-values of `shapiro.test()` and `bartlett.test()` below are both greater than 0.05, indicating the null hypotheses that each group is normally distributed and the variances of groups are similar cannot be rejected. The homogeneity of variances of the dependent variable biomass is also illustrated in the figure.

```
## $`10`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.98388, p-value = 0.9244
##
##
## $`15`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.78665, p-value = 0.08036
##
##
## $`20`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
```

```
## W = 0.86994, p-value = 0.2975
##
##
## $`25`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.86927, p-value = 0.2948
##
##
## $`30`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97163, p-value = 0.8516
##
##
## $`35`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95198, p-value = 0.7285
##
##  Bartlett test of homogeneity of variances
##
## data:  biomass by factor(salt)
## Bartlett's K-squared = 4.5717, df = 5, p-value = 0.4703
```



Problem 3

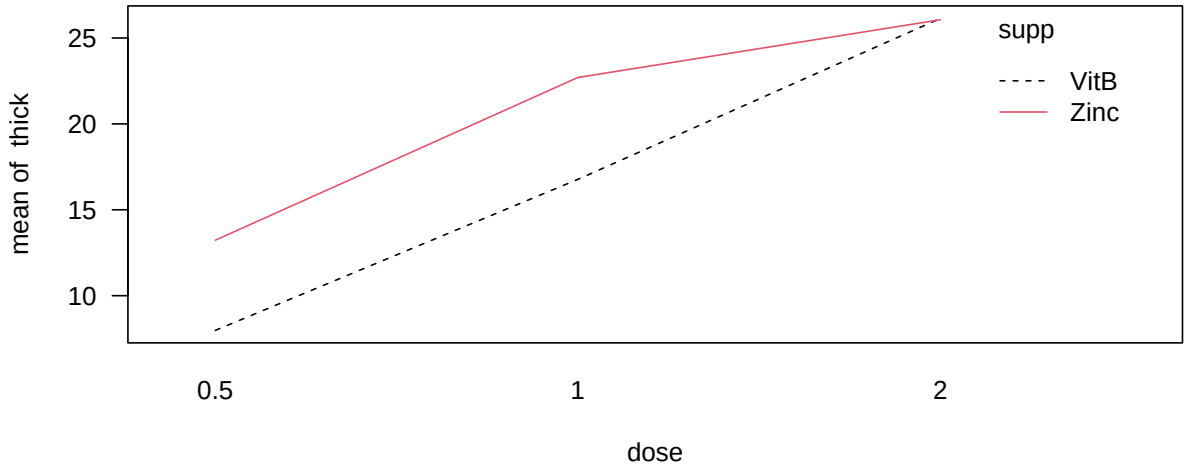
Hypothesis

The null hypothesis is that there is no significant effect of the levels of dose, thickness, and their interactions on the thickness of pangolin scales. While the alternative hypothesis is that there is a significant effect of dose and supplement on the thickness of pangolin scales, and there is an interaction between the two factors.

Model

A two-way ANOVA is appropriate for this study because there are two categorical independent variables (dose and supplement) and one dependent variable (scale thickness). The number of subjects for the two variables is both less than 6, thus, the random effect will not be considered.

In contrast to problem 2, the effects of both supplement and dose are important in the hypothesis being tested in this scenario. Therefore, it is necessary to consider the interaction between these two factors in the model construction. Additionally, the non-parallel lines in the interaction plots further highlight the need to include an interaction term in the model.




```
##
## Call:
## lm(formula = thick ~ factor(dose) * factor(supp), data = scale)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       7.980      1.148   6.949 4.98e-09 ***
## factor(dose)1                     8.790      1.624   5.413 1.46e-06 ***
## factor(dose)2                    18.160      1.624  11.182 1.13e-15 ***
## factor(supp)Zinc                  5.250      1.624   3.233  0.00209 **
## factor(dose)1:factor(supp)Zinc    0.680      2.297   0.296  0.76831
## factor(dose)2:factor(supp)Zinc   -5.330      2.297  -2.321  0.02411 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared:  0.7937, Adjusted R-squared:  0.7746
## F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

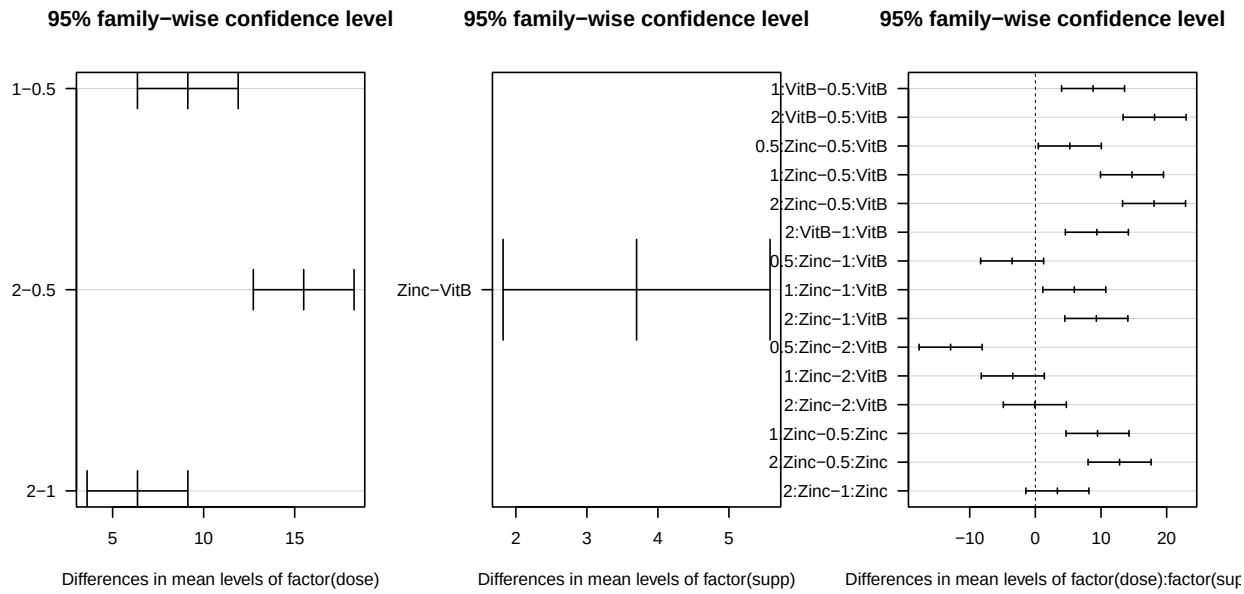
Similarly, the results show that the model is significant as a whole (R^2 = 0.775, $F_{5,54}$ = 41.56, p = < 2.2e-16). For the hypothesis, the statistical results suggest that both dose and supplement have a significant effect on the thickness of pangolin scales with p-values less than 0.05. The reference situation reflected by the intercept is the thickness of scales for pangolins (7.98) receiving 0.5 doses of Vitamin B, which is significantly different from 0 according to the p-value. And the slope coefficients of all the individual dose and supplement factors are positive, indicating the positive relationship between the thickness and the amount of the supplement received in whatever types of nutrient (Vitamin B or Zinc)

Moreover, there is also a significant interaction effect between the two factors, as indicated by the p-value associated with one of the interaction terms. The slope coefficient for the interaction term "dose 2:supplement Zinc" is -5.33, indicating that one unit change in dose 2 (from level 1 to level 2) will cause the impact of changing supplement VitB to ZinC on the mean thickness to decrease by 5.33, and vice versa. The slope

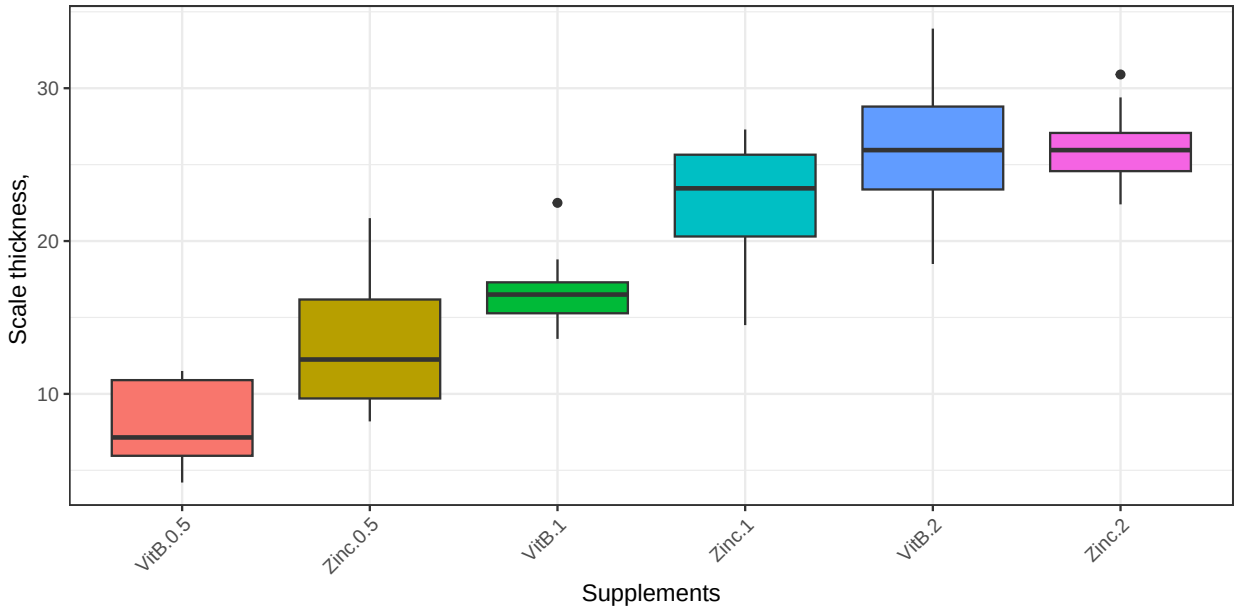
coefficient of interaction term actually reports how a change in one variable impacts its interacted variable's effect on the dependent variable.

Post-hoc test

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $`factor(dose)`
##         diff       lwr       upr   p adj
## 1-0.5  9.130  6.362488 11.897512 0.0e+00
## 2-0.5 15.495 12.727488 18.262512 0.0e+00
## 2-1    6.365  3.597488  9.132512 2.7e-06
##
## $`factor(supp)`
##           diff      lwr      upr     p adj
## Zinc-VitB  3.7 1.820172 5.579828 0.0002312
##
## $`factor(dose):factor(supp)`
##                       diff         lwr       upr     p adj
## 1:VitB-0.5:VitB     8.79   3.9918762 13.588124 0.0000210
## 2:VitB-0.5:VitB    18.16  13.3618762 22.958124 0.0000000
## 0.5:Zinc-0.5:VitB   5.25   0.4518762 10.048124 0.0242521
## 1:Zinc-0.5:VitB    14.72   9.9218762 19.518124 0.0000000
## 2:Zinc-0.5:VitB    18.08  13.2818762 22.878124 0.0000000
## 2:VitB-1:VitB       9.37   4.5718762 14.168124 0.0000058
## 0.5:Zinc-1:VitB    -3.54  -8.3381238  1.258124 0.2640208
## 1:Zinc-1:VitB       5.93   1.1318762 10.728124 0.0073930
## 2:Zinc-1:VitB       9.29   4.4918762 14.088124 0.0000069
## 0.5:Zinc-2:VitB   -12.91 -17.7081238 -8.111876 0.0000000
## 1:Zinc-2:VitB      -3.44  -8.2381238  1.358124 0.2936430
## 2:Zinc-2:VitB      -0.08  -4.8781238  4.718124 1.0000000
## 1:Zinc-0.5:Zinc     9.47   4.6718762 14.268124 0.0000046
## 2:Zinc-0.5:Zinc    12.83   8.0318762 17.628124 0.0000000
## 2:Zinc-1:Zinc       3.36  -1.4381238  8.158124 0.3187361
```

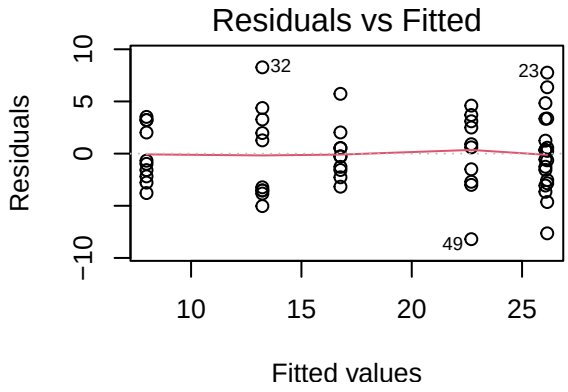


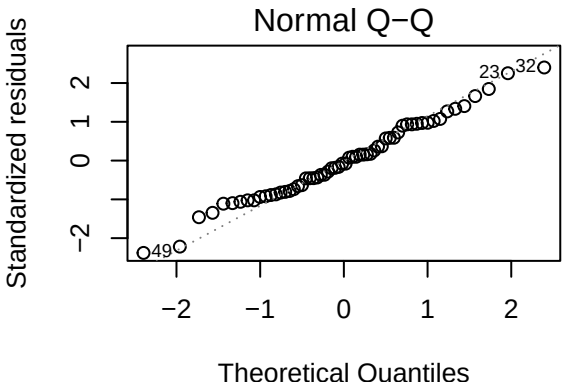
From the Tukey tests, we can find significant differences between all the different dose and supplement groups. And for the interaction groups, the situations are more complicated and I have plotted another graph to better visualize that. We can see that more than a half interaction groups are significantly different while a few are similar.

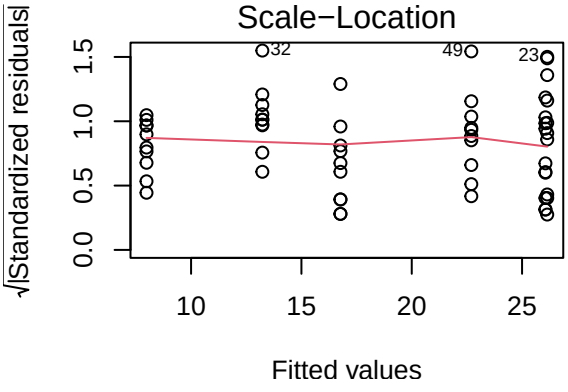


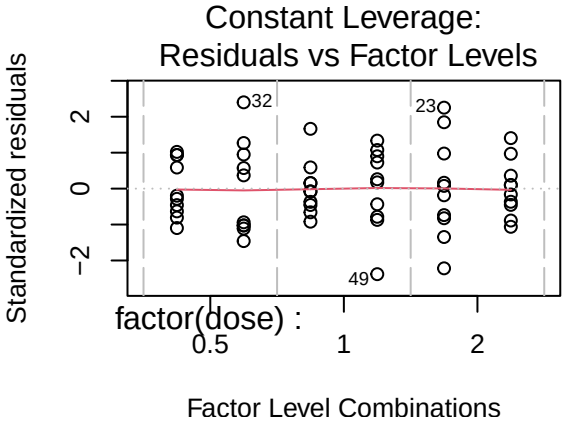
Assumptions

The diagnosis plots for this model residuals are highly satisfactory. We can find a clear stright line in qqplot and residuals plots. Besides, the residuals are equally and randomly distributed in the first and third graphs. Moreover, the leverage plot also does not report any noticeable outliers.


Residuals vs Fitted
Residuals
Fitted values
32
23
49
10
15
20
25
−10
0
5
10


Normal Q−Q
Standardized residuals
Theoretical Quantiles
23
32
49
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
Fitted values
32
49
23
10
15
20
25
0.0
0.5
1.0
1.5


Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
Factor Level Combinations
factor(dose) :
32
23
49
0.5
1
2
−2
0
2

```r
knitr::opts_chunk$set(echo = FALSE, eval = TRUE, warning = F, message = FALSE)
pacman::p_load(ggplot2, mosaic, agricolae)
Tusk <- read.csv("./labs/lab7 linear models with nominal explanatory variables/TuskData.csv")

lm1 <- lm(Tusk.kg ~ factor(Year), data = Tusk)
summary(lm1)
TukeyHSD(lm1)
plot(TukeyHSD(lm1))
# Calculate mean and standard error for each year
mean_data <- aggregate(Tusk$Tusk.kg, by=list(Year=Tusk$Year), FUN=mean)
names(mean_data)[2] <- "Mean"
se_data <- aggregate(Tusk$Tusk.kg, by=list(Year=Tusk$Year), FUN=sd)
se_data$se <- se_data$x / sqrt(length(Tusk$Tusk.kg))

# Create bar plot with error bars
ggplot(mean_data, aes(x=Year, y=Mean, fill=Year)) +
  geom_bar(stat="identity", position="dodge", color="black") +
  geom_errorbar(aes(ymin=Mean-se_data$se, ymax=Mean+se_data$se), width=5) +
  labs(title="Average Weight of Confiscated Elephant Tusks by Year", x="Year", y="Weight (kg)") +
  theme_bw()
par(mfrow=c(2,2), mar = c(3.8, 4, 3, 2))
plot(lm1)
a = with(Tusk, sd(Tusk.kg[Year == 1970])/sd(Tusk.kg[Year == 1990]))
b = with(Tusk, sd(Tusk.kg[Year == 1990])/sd(Tusk.kg[Year == 2010]))
c = with(Tusk, sd(Tusk.kg[Year == 1970])/sd(Tusk.kg[Year == 2010]))
print(paste("The ratios of sample dependent variables' sds are", a,",", b,",", c))
salt <- read.csv("./labs/lab7 linear models with nominal explanatory variables/salt.csv")
# salt <- read.csv("salt.csv")
# salt$inter <- interaction(salt$salt, salt$block)
with(salt, interaction.plot(salt, block, biomass, col = c(1,
2, 3, 4), las = 1, cex = 0.9))
lm2 <- lm(biomass ~ factor(salt) + factor(block), data = salt)
summary(lm2)
TukeyHSD(lm2)

par(mfrow=c(1,2))
plot(TukeyHSD(lm2), cex.axis=0.5, las = 1)
par(mfrow=c(2,2), mar = c(3.8, 4, 3, 2))
plot(lm2)
groups <- split(salt$biomass, salt$salt)
lapply(groups, shapiro.test)

# var.test(biomass ~ salt, data = salt)
bartlett.test(biomass~factor(salt), data = salt)

require(graphics)
plot(biomass ~ factor(salt), data = salt)
scale <- read.csv("./labs/lab7 linear models with nominal explanatory variables/ScaleThickness.csv")
with(scale, interaction.plot(dose, supp, thick, col = c(1,
2, 3, 4), las = 1, cex = 0.9))
lm3 <- lm(thick ~ factor(dose) * factor(supp), data = scale)
summary(lm3)
TukeyHSD(lm3)
```

```r
par(mfrow=c(1,3))
plot(TukeyHSD(lm3), las =1)
scale$inter <- interaction(scale$supp, scale$dose)
ggplot(data = scale, aes(y = thick, x = inter)) +
  geom_boxplot(aes(fill = inter)) +
  labs(y = "Scale thickness, ", x = "Supplements") + theme_bw() +
  theme(legend.position = "none",
      axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
par(mfrow=c(2,2), mar = c(3.8, 4, 3, 2))
plot(lm3)
```