ENV 710: Lab 7

Jiahuan Li

Spring 2023

```
library(ggplot2)
require(mosaic)
```

Problem 1

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.

To test this hypothesis, we will use a one-way ANOVA design with the year as a factor because there is only one independent variable for this model. Besides, the number of the subjects in this situation is too small to conduct a random effect analysis. Thus, I construct the model as follows:

```
Tusk <- read.csv("TuskData.csv")

lm1 <- lm(Tusk.kg ~ factor(Year), data = Tusk)
summary(lm1)
```

```
##
## 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
```

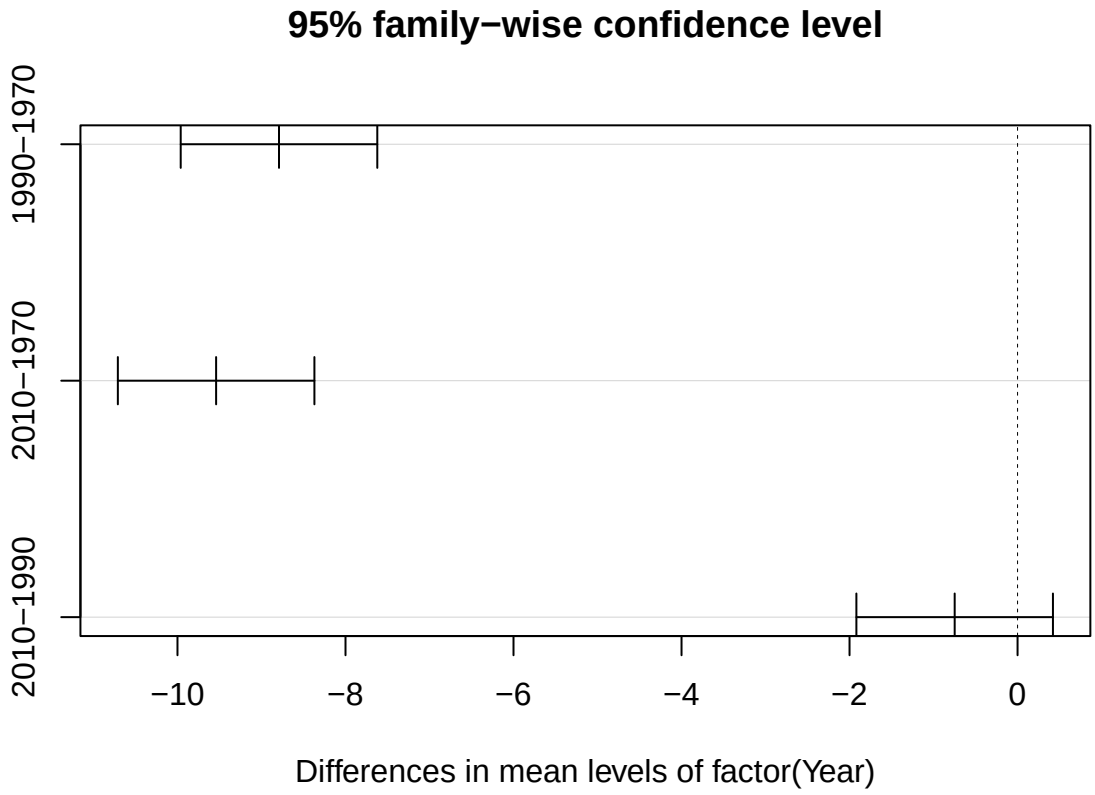
The ANOVA results show that there is a statistically significant difference in the mean weight of elephant tusks across the three years ($F(2, 57) = 238.1$, $p < 2.2e\text{-}16$). The mean weight of tusks appears to have decreased over time, with the lowest average weight in 2010.

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

```
TukeyHSD(lm1)
```

```
##   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
```

```
plot(TukeyHSD(lm1))
```

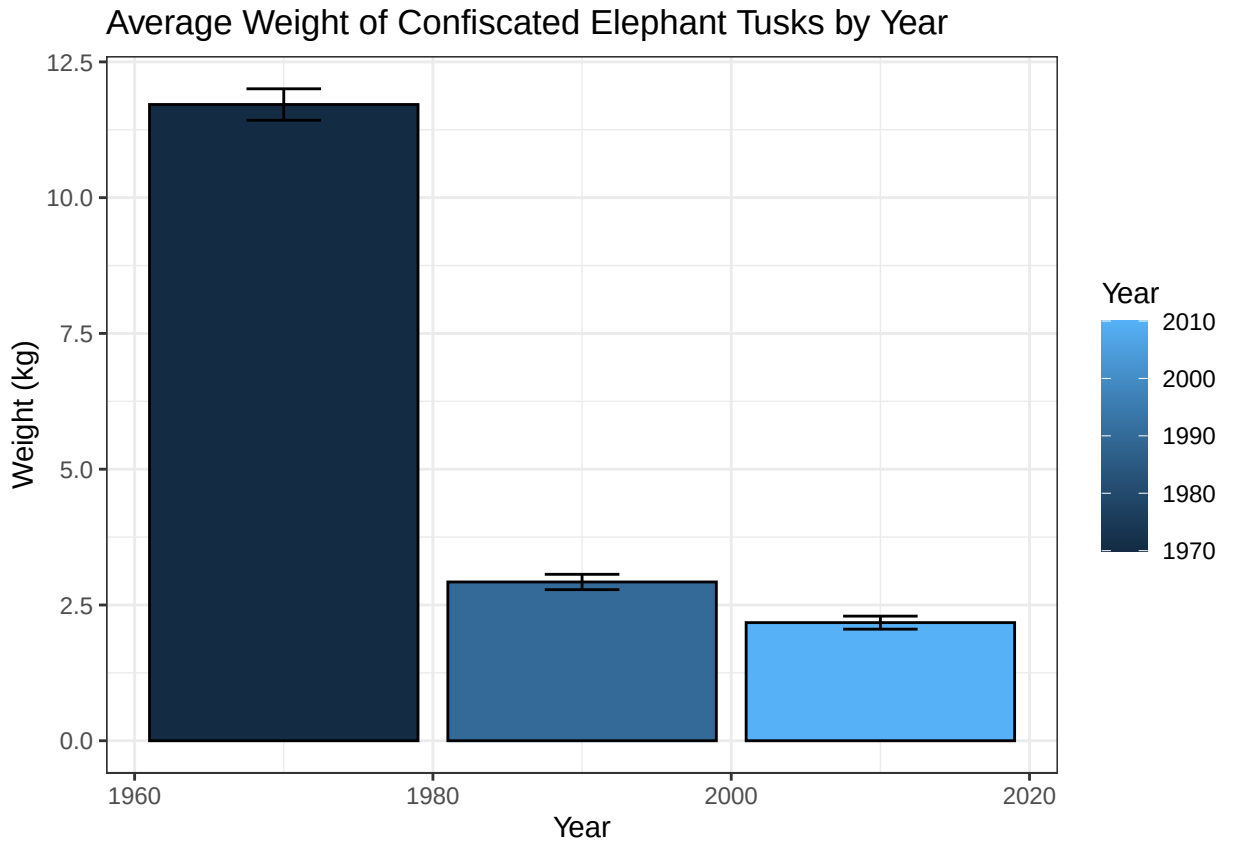


In this example, there is not a statistically significant difference between year 2010-1990 as illustrated by the fact that the CI overlaps 0 and the p-value is greater than 0.05. By contrast, there appears to be significant differences in biomass between year 2010-1970 and year 1990-1970.

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.

```r
# 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()
```

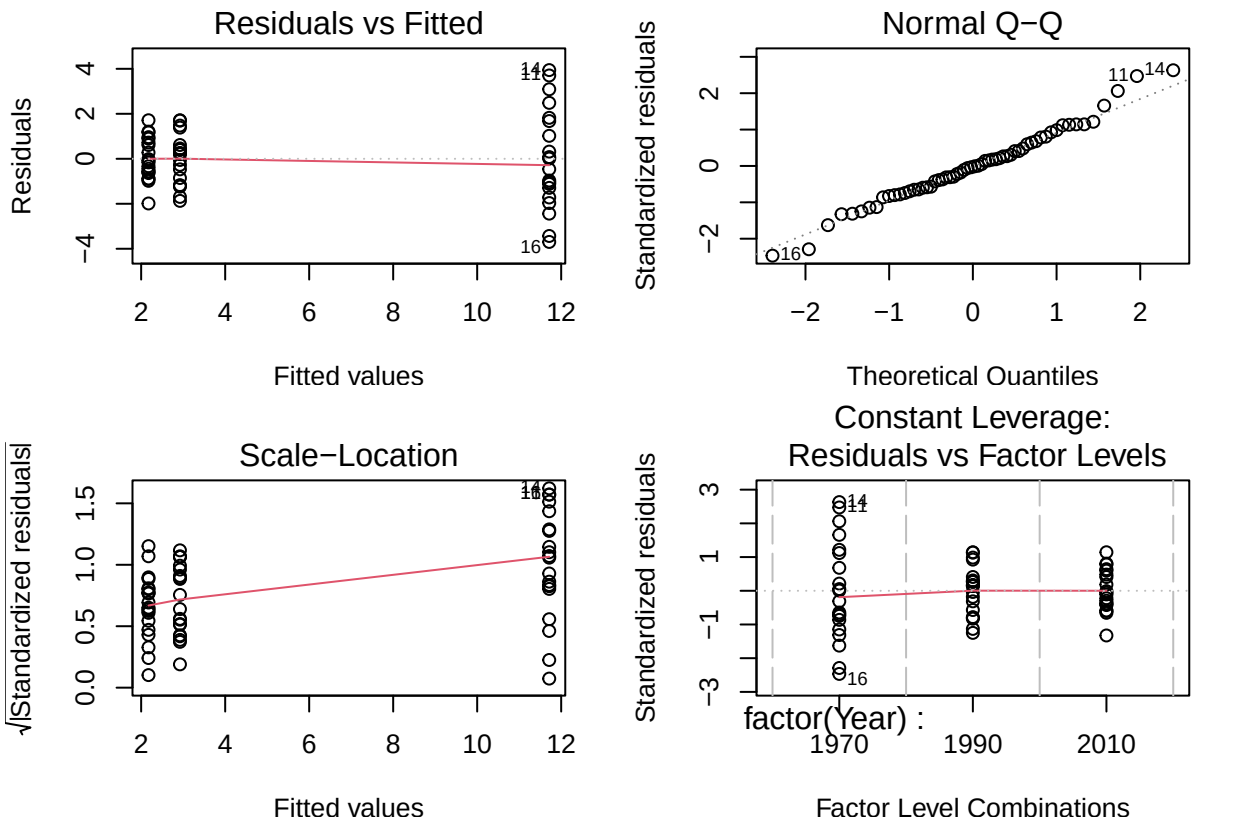


Then, I checked the assumptions of the statistical test using diagnosis plots. First, In the residual plot, the residuals randomly scattered around the centerline, indicating residuals roughly normally and approximately independently distributed with a mean of 0 and constant variance. Second, the qq plot shows a linear trend, indicating the normality of the dependent variables is good.

Third, however, the figure of standardized residual plot is worrying since the variance seems to increase with the treatment means, which might be evidence of heteroscedasticity. Thus, I double checked the ratios of the sample standard deviations and find some of them exceed the limit of 2. To remedy this, I transform the dependent variable into a log format and rerun the model. But the ratio also cannot meet the requirement. Thus, I decide to ignore this problem since other plots are good.

At last, this plot, the leverage plot does not reflect any outliers, although observations 14 and 16 have more leverage than other observations.

```
par(mfrow=c(2,2), mar = c(3.8, 4, 3, 2))
plot(lm1)
```




```
with(Tusk, sd(Tusk.kg[Year == 1970])/sd(Tusk.kg[Year == 1990]))
```

```
## [1] 2.051206
```

```
with(Tusk, sd(Tusk.kg[Year == 1990])/sd(Tusk.kg[Year == 2010]))
```

```
## [1] 1.179337
```

```
with(Tusk, sd(Tusk.kg[Year == 1970])/sd(Tusk.kg[Year == 2010]))
```

```
## [1] 2.419062
```

Problem 2

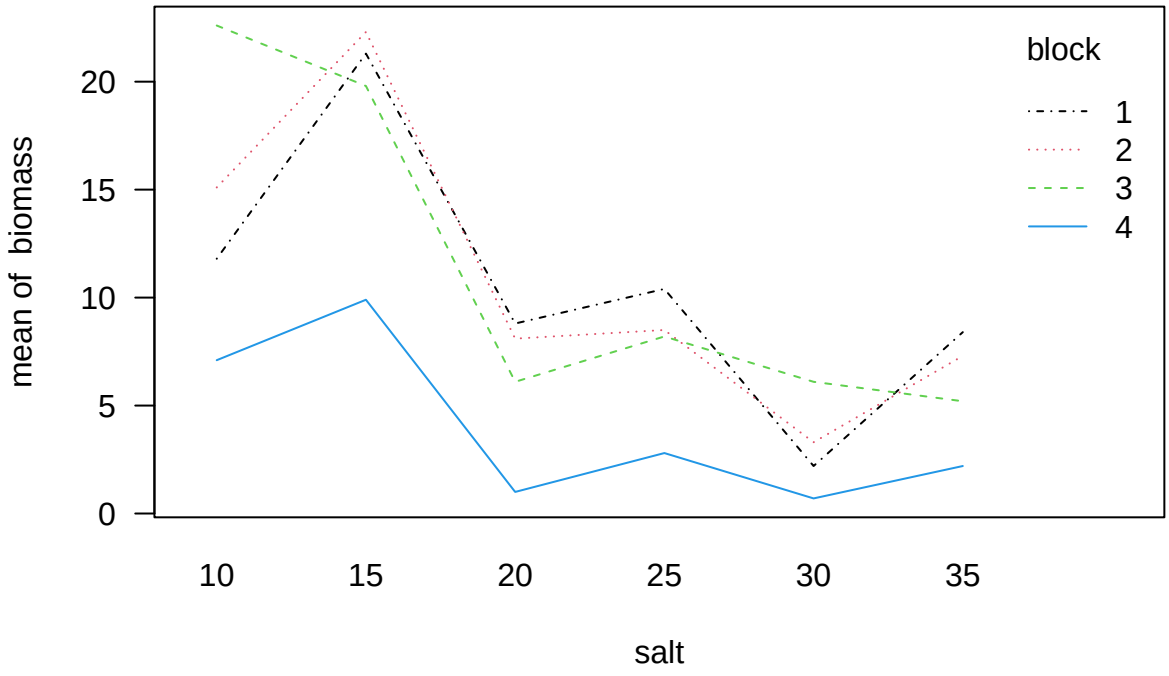
The null hypothesis is that the mean biomass growth is the same across all six levels of salt addition. And the alternative hypothesis is that the mean biomass growth is different across at least one pair of salt addition levels.

To test this hypothesis, I use a fixed complete block design with salt addition as the factor and block as the blocking variable. That is because the assignment of treatments is carried out separately within each

block with every treatment included at least once in every block. Besides, the number of the subjects in this situation is 4, which is too small to conduct a random effect analysis.

Moreover, I have made interaction plots to determine whether an interaction term should be included in the model. The lines shown in the plot are nearly parallel to each other, indicating that the model of interaction.

```
# 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))
```



Thus, I construct the model as follows:

```
lm2 <- lm(biomass ~ factor(salt) + factor(block), data = salt)

TukeyHSD(lm2)
```

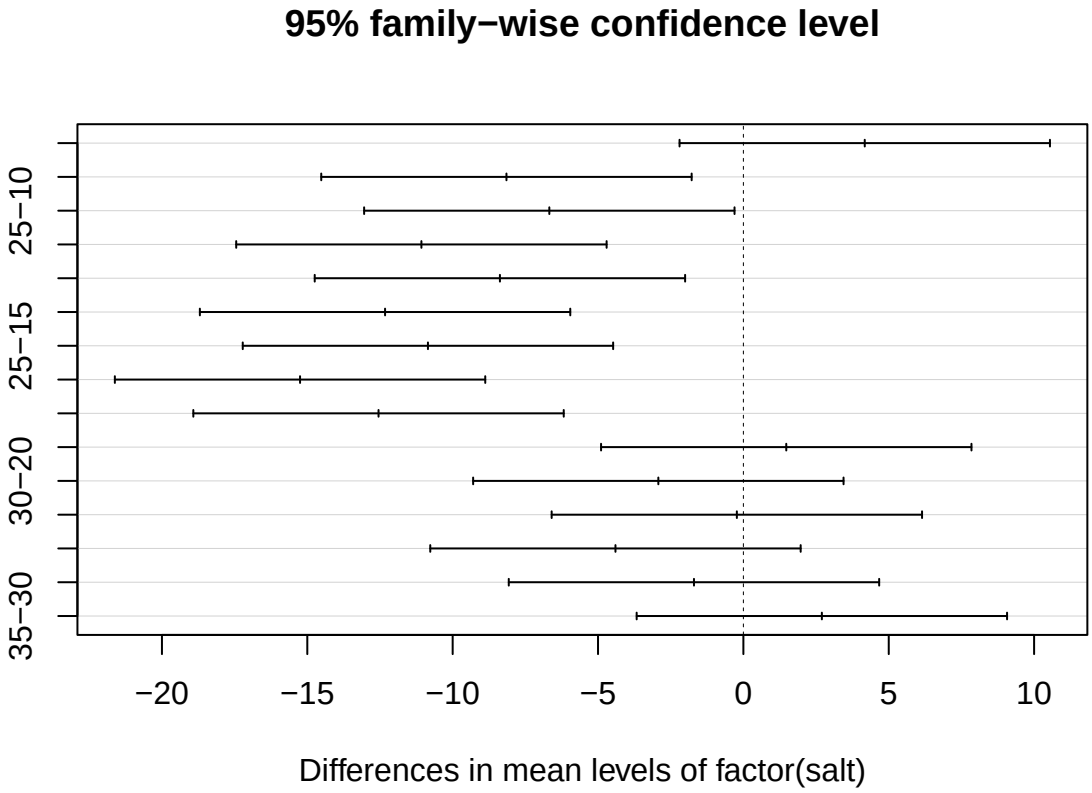
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $`factor(salt)`
##          diff        lwr        upr     p adj
## 15-10   4.175  -2.195427 10.5454267 0.3242073
```

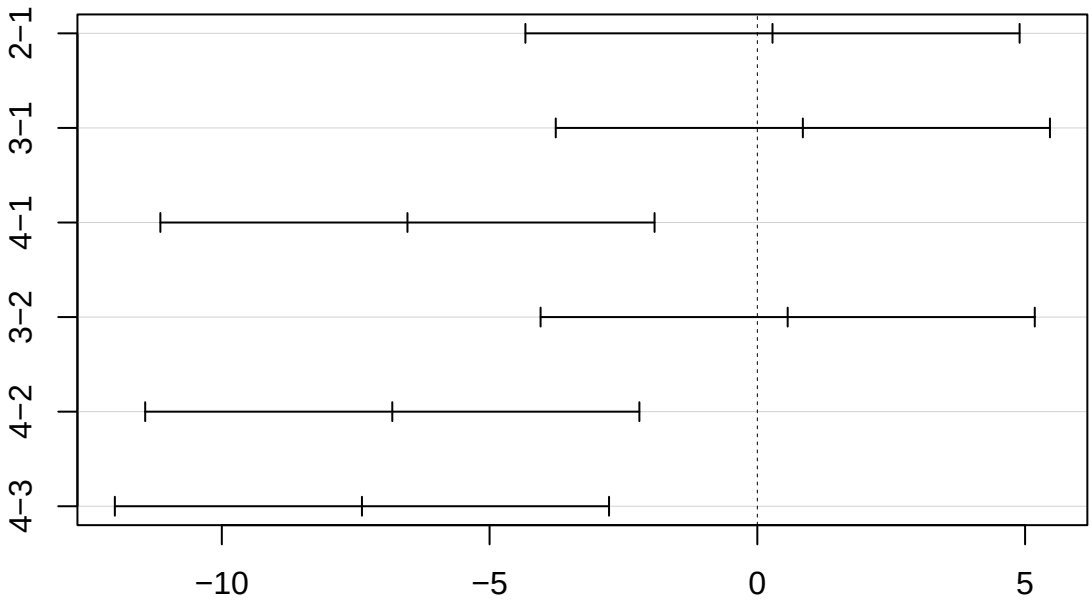
```
## 20-10  -8.150 -14.520427 -1.7795733 0.0089392
## 25-10  -6.675 -13.045427 -0.3045733 0.0374481
## 30-10 -11.075 -17.445427 -4.7045733 0.0005436
## 35-10  -8.375 -14.745427 -2.0045733 0.0071740
## 20-15 -12.325 -18.695427 -5.9545733 0.0001755
## 25-15 -10.850 -17.220427 -4.4795733 0.0006697
## 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
```

```
summary(lm2)
```

```
##
## 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
```

```
plot(TukeyHSD(lm2))
```






```
ad <- read.csv("Sales.csv", header = T)
ad <- with(ad, data.frame(sales, city = factor(city), campaign = factor(campaign),
time = factor(time)))

mod.sales <- lm(sales ~ campaign * time, data = ad)
summary(mod.sales)
```

```
##
## Call:
## lm(formula = sales ~ campaign * time, data = ad)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -378.4 -199.0   26.6  241.9  317.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        718.6      123.8   5.804 5.53e-06 ***
## campaign2         -140.4      175.1  -0.802    0.430
## time2               85.6      175.1   0.489    0.629
## time3              -23.2      175.1  -0.133    0.896
## campaign2:time2    -10.4      247.6  -0.042    0.967
## campaign2:time3    -17.6      247.6  -0.071    0.944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 276.8 on 24 degrees of freedom
## Multiple R-squared:  0.1135, Adjusted R-squared:  -0.07113
## F-statistic: 0.6148 on 5 and 24 DF,  p-value: 0.6896
```

```
require(lme4)
```

```
## Loading required package: lme4
```

```
##
## Attaching package: 'lme4'
```

```
## The following object is masked from 'package:mosaic':
##
##     factorize
```

```
require(lmerTest)
```

```
## Loading required package: lmerTest
```

```
## Warning: package 'lmerTest' was built under R version 4.2.3
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##     lmer
```

```
## The following object is masked from 'package:mosaic':
##
##     rand
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
with(ad, tapply(sales, list(time), mean))
```

```
##     1     2     3
## 648.4 728.8 616.4
```

```
with(ad, tapply(sales, list(campaign), mean))
```

```
##        1        2
## 739.4000 589.6667
```

```r
mod.rep <- lmer(sales ~ campaign + time + (1 | city/campaign),
data = ad)
summary(mod.rep)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: sales ~ campaign + time + (1 | city/campaign)
##    Data: ad
##
## REML criterion at convergence: 286.4
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.72966 -0.41252  0.08216  0.58379  1.88255
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  campaign:city (Intercept) 75045.0  273.94
##  city          (Intercept)  1245.7   35.29
##  Residual                    339.9   18.44
## Number of obs: 30, groups:  campaign:city, 10; city, 10
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  723.267    123.707    8.024   5.847  0.00038 ***
## campaign2   -149.733    174.819    8.000  -0.857  0.41663
## time2         80.400      8.245   18.000   9.751 1.32e-08 ***
## time3        -32.000      8.245   18.000  -3.881  0.00110 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) cmpgn2 time2
## campaign2 -0.707
## time2     -0.033  0.000
## time3     -0.033  0.000  0.500
```