STAT461 HW7

Xiangyu Ren

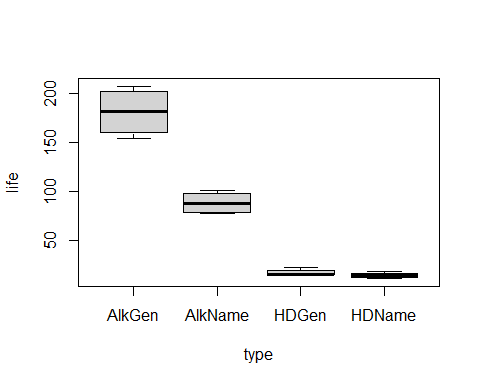
10/23/2020

#### Problem 1. Your homework will consider an experiment on battery life for different types and brands of battery. Two brands (a name brand and a generic brand) of two types (Alkaline and “Heavy Duty”) of batteries were tested to see how long they could run continuously. This results in four categories, AlkName is for name-brand alkaline batteries, AlkGen is for generic alkaline batteries, HDName is for heavy duty name-brand batteries, and HDGen is for generic heavy duty batteries. Four batteries of each type were tested and the times to battery failure are recorded as below. Use the code below to read in the data:

type<-c("AlkName","AlkName","AlkName","AlkName","AlkGen","AlkGen","AlkGen","AlkGen",  
"HDName","HDName","HDName","HDName","HDGen","HDGen","HDGen","HDGen")  
life<-c(100.668, 77.734,79.210,95.063,206.880,153.347,165.980,196.000,  
14.951,18.063,11.111,12.840,15.340,22.090,15.734, 14.440)  
batt<-data.frame(type=type, life=life)

#### 1. Plot the data.

boxplot(life ~ type, data = batt)



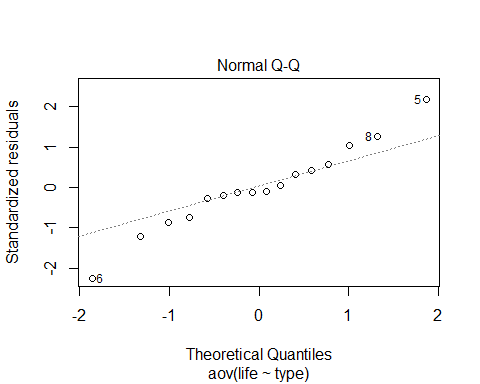
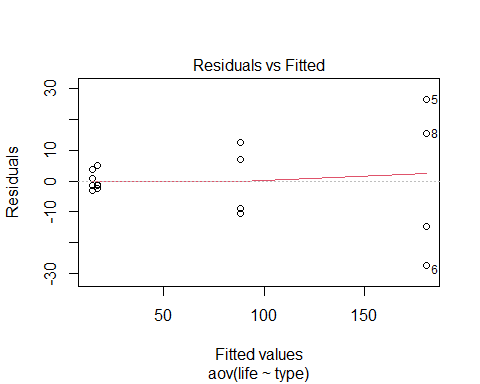
#### 2. For the battery data, do the following:

#### (a) Write out the one-way ANOVA model for this data.

,

#### Show residual plots for this model. Are the residuals approximately normal? Justify your answer.

model1 = aov(life ~ type, data = batt)  
plot(model1, which = c(1, 2))



In the first plot, it seems like the abline is a straight line, and in the QQ plot, all the observations are like having linear relation, thus we say the residuals are approximately normal.

#### (c) Is the assumption of constant error variance among treatments justified? Explain your answer.

v1 = var(batt$life[batt$type == "AlkName"])  
v2 = var(batt$life[batt$type == "AlkGen"])  
v3 = var(batt$life[batt$type == "HDName"])  
v4 = var(batt$life[batt$type == "HDGen"])  
v1;v2;v3;v4

## [1] 130.9684

## [1] 628.0865

## [1] 8.957162

## [1] 12.26028

v2/v3

## [1] 70.12115

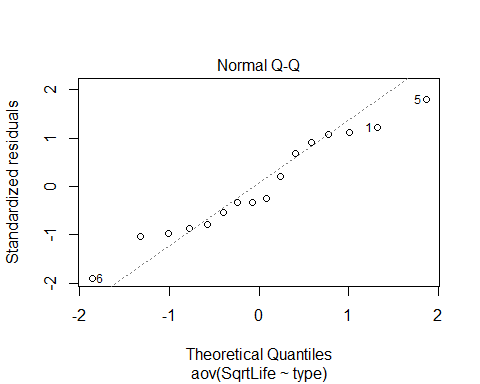
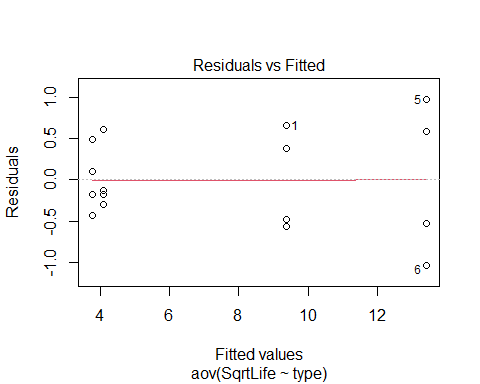
The assumption of constant error variance among treatment is not justified. Because the largest variance is much larger than the smallest one, .

#### 3. Now consider using the square-root of the battery life as a response variable. Repeat (a)-(c) above for this transformation.

batt$SqrtLife = sqrt(life)  
batt$LogLife = log(life)  
batt$SqrLife = (life)^2

,

model.sqrt = aov(SqrtLife ~ type, data = batt)  
plot(model.sqrt, which = c(1, 2))



In the first plot, it seems like the abline is a straight line, and in the QQ plot, all the observations are like having linear relation, thus we say the residuals are approximately normal.

v1 = var(batt$SqrtLife[batt$type == "AlkName"])  
v2 = var(batt$SqrtLife[batt$type == "AlkGen"])  
v3 = var(batt$SqrtLife[batt$type == "HDName"])  
v4 = var(batt$SqrtLife[batt$type == "HDGen"])  
v1;v2;v3;v4

## [1] 0.370466

## [1] 0.875621

## [1] 0.1549349

## [1] 0.1671155

v2/v3

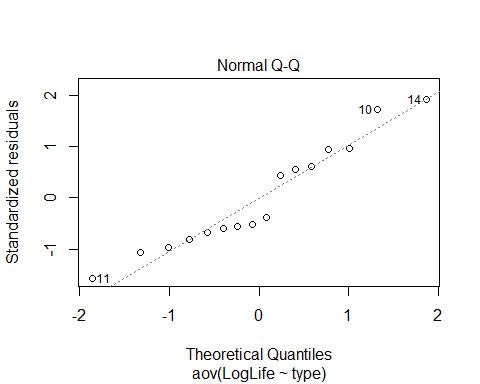
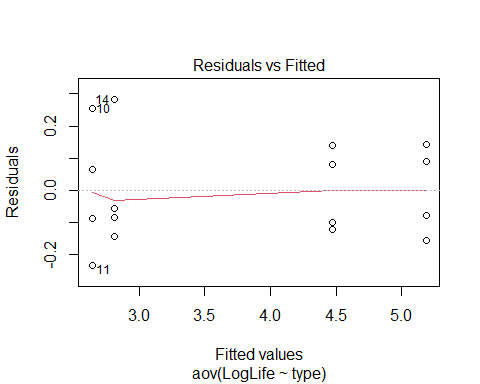
## [1] 5.651542

The assumption of constant error variance among treatment is not justified. Because the largest variance is much larger than the smallest one, .

#### 4. Now consider using the log of the battery life as a response variable. Repeat (a)-(c) above for this transformation.

,

model.log = aov(LogLife ~ type,data = batt)  
plot(model.log, which = c(1,2))



In the first plot, it seems like the abline is a straight line, and in the QQ plot, all the observations are like having linear relation, thus we say the residuals are approximately normal.

v1 = var(batt$LogLife[batt$type == "AlkName"])  
v2 = var(batt$LogLife[batt$type == "AlkGen"])  
v3 = var(batt$LogLife[batt$type == "HDName"])  
v4 = var(batt$LogLife[batt$type == "HDGen"])  
v1;v2;v3;v4

## [1] 0.01681069

## [1] 0.01960264

## [1] 0.04338084

## [1] 0.03670805

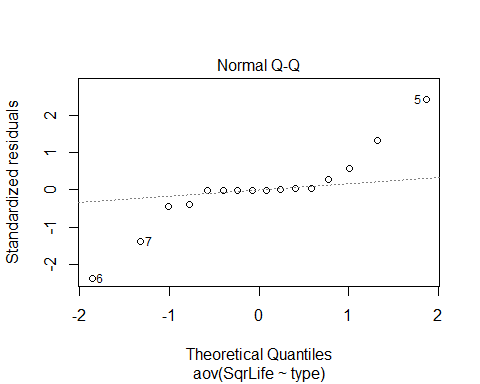
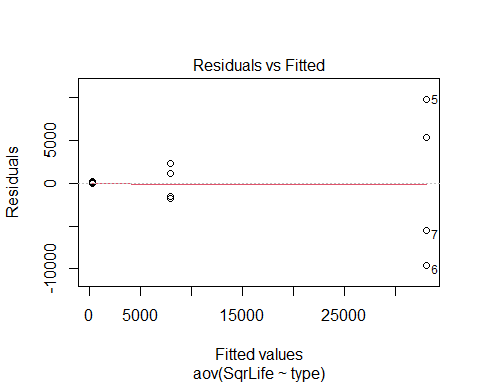
v3/v1

## [1] 2.580551

The assumption of constant error variance among treatment is justified. Because the largest variance is not much larger than the smallest one, .

#### 5. Now consider using the square of the battery life as a response variable. Repeat (a)-(c) above for this transformation.

model.sqr = aov(SqrLife ~ type, data = batt)  
plot(model.sqr, which = c(1, 2))



We can see clearly that in the QQ plot, the dots are not one a straight line, so the residuals are not normal distributed.

v1 = var(batt$SqrLife[batt$type == "AlkName"])  
v2 = var(batt$SqrLife[batt$type == "AlkGen"])  
v3 = var(batt$SqrLife[batt$type == "HDName"])  
v4 = var(batt$SqrLife[batt$type == "HDGen"])  
v1;v2;v3;v4

## [1] 4124567

## [1] 81669791

## [1] 7742.895

## [1] 16843.13

v2/v3

## [1] 10547.71

The assumption of constant error variance among treatment is not justified. Because the largest variance is much larger than the smallest one, .

#### 6. Which of the four models you have fit has residuals that best satisfy the assumptions of the ANOVA model? Explain your choice.

summary(model1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## type 3 73527 24509 125.6 2.5e-09 \*\*\*  
## Residuals 12 2341 195   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(model.sqrt)

## Df Sum Sq Mean Sq F value Pr(>F)   
## type 3 255.8 85.28 217.5 1.01e-10 \*\*\*  
## Residuals 12 4.7 0.39   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(model.log)

## Df Sum Sq Mean Sq F value Pr(>F)   
## type 3 18.80 6.267 215.2 1.08e-10 \*\*\*  
## Residuals 12 0.35 0.029   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(model.sqr)

## Df Sum Sq Mean Sq F value Pr(>F)   
## type 3 2.905e+09 968365217 45.13 8.23e-07 \*\*\*  
## Residuals 12 2.575e+08 21454736   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

I think the log transformation best fit the residuals to satisfy teh assumptions of the anova model, because, we can see from the outputs taht the log model has the lowest Mean Sq, and also, log model is the only model that justified the assumption of constant error variance among treatment.

#### 7. For the model you chose in question 6 above, are there any significant pairwise differences in mean lifetime of different battery types? If so, state which are different, and provide p-values, test statistics, and null hypotheses for the hypothesis tests used.

We give the hypothesis test with following

library(lsmeans)

## Loading required package: emmeans

## The 'lsmeans' package is now basically a front end for 'emmeans'.  
## Users are encouraged to switch the rest of the way.  
## See help('transition') for more information, including how to  
## convert old 'lsmeans' objects and scripts to work with 'emmeans'.

lsm.batt = lsmeans(model.log, "type")  
library(knitr)  
summary(contrast(lsm.batt, method = "pairwise", adjust = "tukey"), infer = c(T, T), level = 0.95, side = "two-sided")

## contrast estimate SE df lower.CL upper.CL t.ratio p.value  
## AlkGen - AlkName 0.716 0.121 12 0.357 1.074 5.931 0.0003   
## AlkGen - HDGen 2.376 0.121 12 2.018 2.734 19.688 <.0001   
## AlkGen - HDName 2.549 0.121 12 2.191 2.907 21.122 <.0001   
## AlkName - HDGen 1.660 0.121 12 1.302 2.018 13.757 <.0001   
## AlkName - HDName 1.833 0.121 12 1.475 2.191 15.191 <.0001   
## HDGen - HDName 0.173 0.121 12 -0.185 0.531 1.434 0.5034   
##   
## 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

We can see that the t ratio for AG-AN is 5.931, for AG-HG is 19.688, for AG-HN is 21.122, for AN-HG is 13.757, for AN-HN is 15.191, for HG-HN is 1.434.

Since the p-value we get is all less than , except for . We can reject the null hypothesis, and conclude that there is at least one different subtraction output.

#### Problem 2. A study was conducted to compare the calories and sodium in hot dogs made with different types of meat. Read the data into R using the following commands:

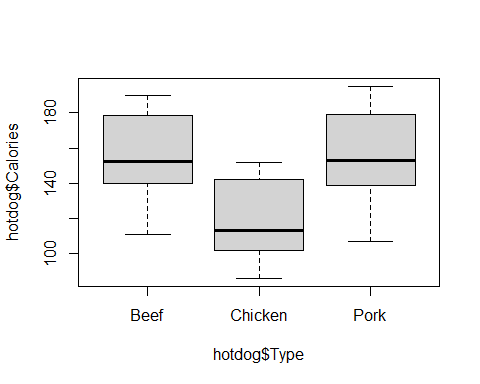
hotdog = read.table("hotdogs.txt", header = TRUE)  
hotdog

## Type Calories Sodium  
## 1 Beef 186 495  
## 2 Beef 181 477  
## 3 Beef 176 425  
## 4 Beef 149 322  
## 5 Beef 184 482  
## 6 Beef 190 587  
## 7 Beef 158 370  
## 8 Beef 139 322  
## 9 Beef 175 479  
## 10 Beef 148 375  
## 11 Beef 152 330  
## 12 Beef 111 300  
## 13 Beef 141 386  
## 14 Beef 153 401  
## 15 Beef 190 645  
## 16 Beef 157 440  
## 17 Beef 131 317  
## 18 Beef 149 319  
## 19 Beef 135 298  
## 20 Beef 132 253  
## 21 Pork 173 458  
## 22 Pork 191 506  
## 23 Pork 182 473  
## 24 Pork 190 545  
## 25 Pork 172 496  
## 26 Pork 147 360  
## 27 Pork 146 387  
## 28 Pork 139 386  
## 29 Pork 175 507  
## 30 Pork 136 393  
## 31 Pork 179 405  
## 32 Pork 153 372  
## 33 Pork 107 144  
## 34 Pork 195 511  
## 35 Pork 135 405  
## 36 Pork 140 428  
## 37 Pork 138 339  
## 38 Chicken 129 430  
## 39 Chicken 132 375  
## 40 Chicken 102 396  
## 41 Chicken 106 383  
## 42 Chicken 94 387  
## 43 Chicken 102 542  
## 44 Chicken 87 359  
## 45 Chicken 99 357  
## 46 Chicken 107 528  
## 47 Chicken 113 513  
## 48 Chicken 135 426  
## 49 Chicken 142 513  
## 50 Chicken 86 358  
## 51 Chicken 143 581  
## 52 Chicken 152 588  
## 53 Chicken 146 522  
## 54 Chicken 144 545

#### and plot calories as a response variable with the type of meat on the x-axis. Your plot could be either a boxplot or a plot with one dot for each hot dog.

#### Answer the following question: “Are there differences in the average calories of hot dogs made with different kinds of meat?”. To answer this question, write down a statistical model (clearly state the response variable, treatment levels, number of replicates, . . . ), express the above question as a testable null hypothesis, and report the p-value of the test statistic under the null hypothesis. Conduct an analysis of pairwise differences if it helps you clarify where there are differences in mean calories. Your answer should include all R code used, and the important R output. If you need to transform the response variable, do so, but you do NOT need to provide details of the transformations that you tried, but did not ultimately select. Only report the best transformation, and only show residual plots, ANOVA tables, and any other results for this transformation.

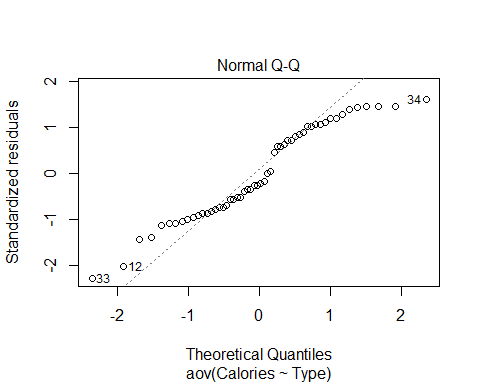
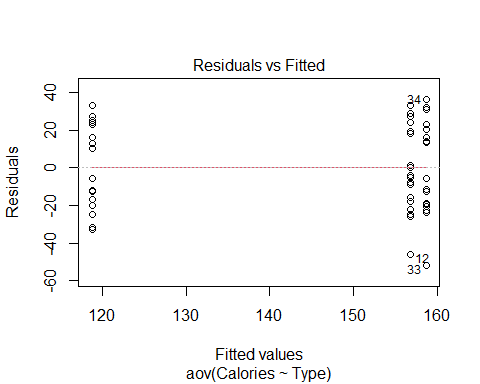
boxplot(hotdog$Calories ~ hotdog$Type, main = "")



,

We give the hypothesis with following

model2 = aov(Calories ~ Type, data = hotdog)  
plot(model2, which = c(1, 2))



v1 = var(hotdog$Calories[hotdog$Type == "Beef"])  
v2 = var(hotdog$Calories[hotdog$Type == "Pork"])  
v3 = var(hotdog$Calories[hotdog$Type == "Chicken"])  
v1;v2;v3

## [1] 512.6605

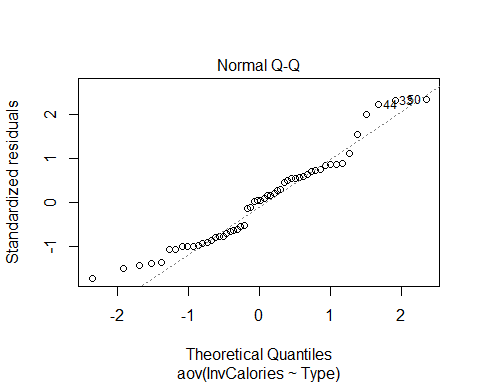
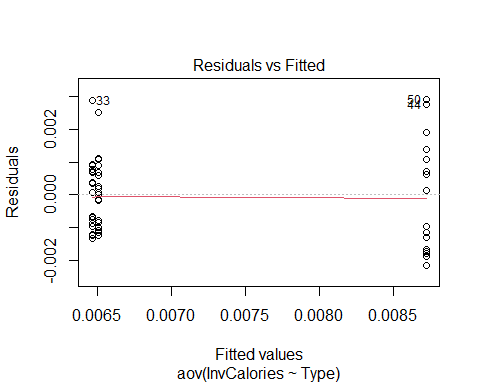
## [1] 636.8456

## [1] 508.5662

v2/v3

## [1] 1.252237

hotdog$InvCalories = 1/hotdog$Calories  
model.inv = aov(InvCalories ~ Type, data = hotdog)  
plot(model.inv, which = c(1, 2))



v1 = var(hotdog$InvCalories[hotdog$Type=="Beef"])  
v2 = var(hotdog$InvCalories[hotdog$Type=="Pork"])  
v3 = var(hotdog$InvCalories[hotdog$Type=="Chicken"])  
v1;v2;v3

## [1] 9.629329e-07

## [1] 1.234792e-06

## [1] 2.875638e-06

v3/v1

## [1] 2.986333

library(lsmeans)  
lsm.Inv=lsmeans(model.inv, "Type")  
library(knitr)  
summary(contrast(lsm.Inv, method = "pairwise", adjust = "tukey"), infer=c(T, T), level = 0.95, side = "two-sided")

## contrast estimate SE df lower.CL upper.CL t.ratio p.value  
## Beef - Chicken -2.21e-03 0.000424 51 -0.003234 -0.00119 -5.223 <.0001   
## Beef - Pork 4.38e-05 0.000424 51 -0.000979 0.00107 0.103 0.9941   
## Chicken - Pork 2.26e-03 0.000440 51 0.001193 0.00332 5.122 <.0001   
##   
## 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

Therefore, as we can get from the outputs by inverse transformation, since all the p-value is smaller than except , we can reject the null hypothesis, and conclude that there is at least 1 different subtraction output.