Assignment 7

Jiarong Ye

October 25, 2018

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,
- 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:

Read Data

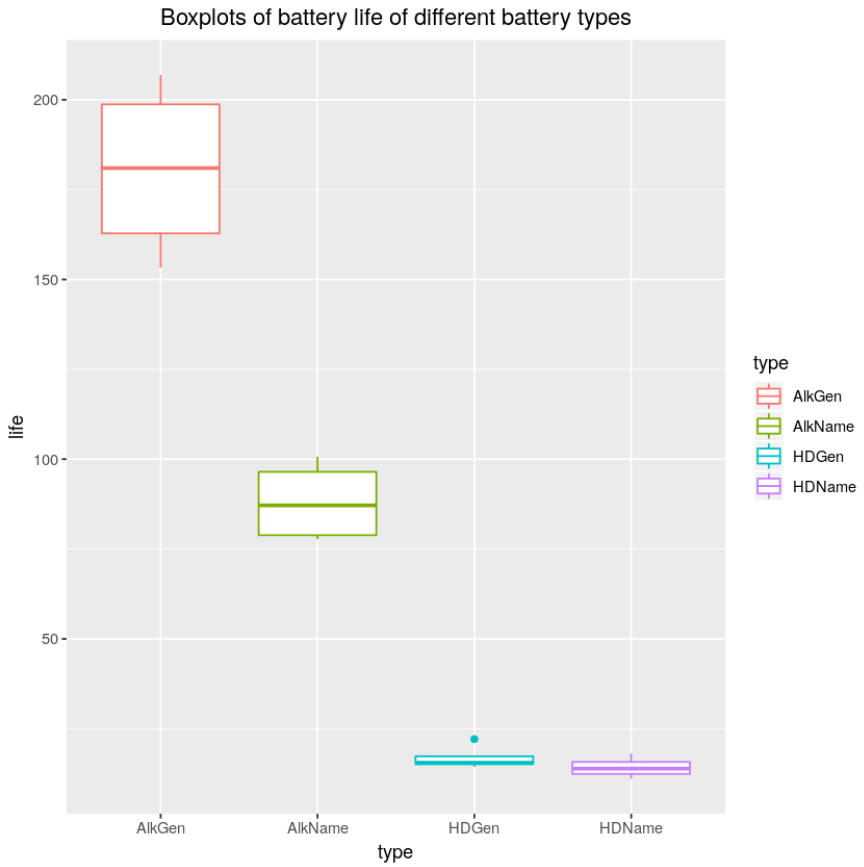
```
In [1]: type=c("AlkName","AlkName","AlkName","AlkName","AlkGen","AlkGen","AlkGen","AlkGen",
        "HDName","HDName","HDName","HDName","HDGen","HDGen","HDGen","HDGen")
        life=c(100.668, 77.734,79.210,95.063,206.880,153.347,165.980,196.000,
        14.951,18.063,11.111,12.840,15.340,22.090,15.734, 14.440)
        batt=data.frame(type=type, life=life)
        batt
```

type	life
AlkName	100.668
AlkName	77.734
AlkName	79.210
AlkName	95.063
AlkGen	206.880
AlkGen	153.347
AlkGen	165.980
AlkGen	196.000
HDName	14.951
HDName	18.063
HDName	11.111
HDName	12.840
HDGen	15.340
HDGen	22.090
HDGen	15.734
HDGen	14.440

Q1

Plot the data

```
In [4]: library(ggplot2)
        ggplot(batt, aes(x=type, y=life, color=type)) +
                geom_boxplot() +
                ylab('life') +
                ggtitle('Boxplots of battery life of different battery types') +
                theme(plot.title = element_text(hjust = 0.5))
```



Q2

For the battery data, do the following:

- (a) Write out the one-way ANOVA model for this data.
- (b) Show residual plots for this model. Are the residuals approximately normal? Justify your answer.
- (c) Is the assumption of constant error variance among treatments justified? Explain your answer.

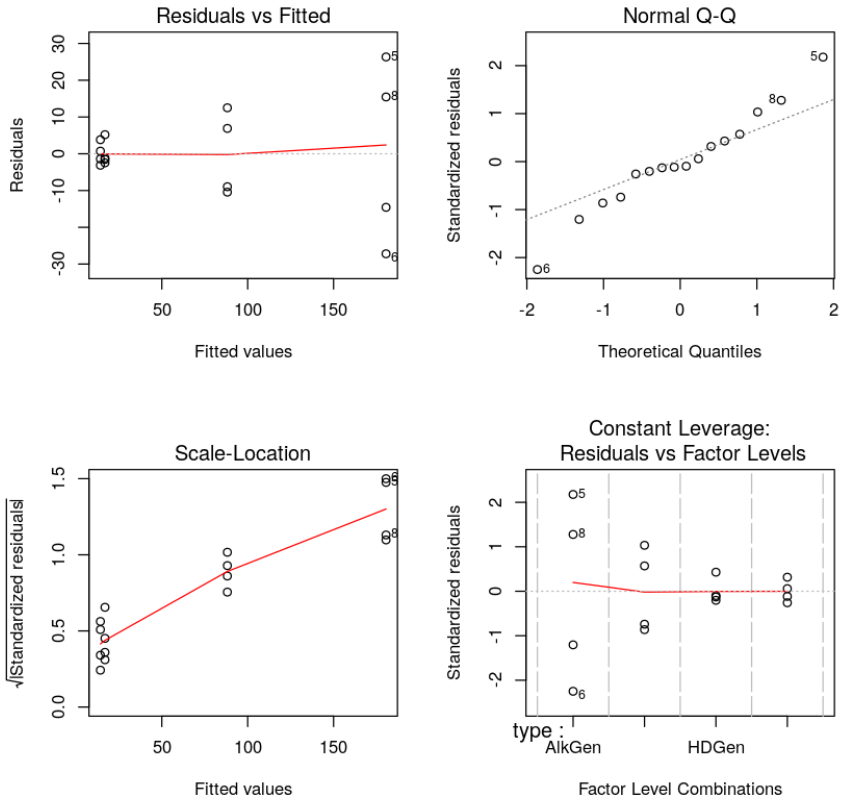
(a)

```
In [7]: library(knitr)
        library(lsmeans)
        aov.batt = aov(life~type, data = batt)
        kable(anova(aov.batt), format='markdown')
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	3	73526.811	24508.9369	125.643	0
Residuals	12	2340.817	195.0681	NA	NA

(b)

```
In [9]: par(mfrow=c(2,2))
        plot(aov.batt)
```



From the QQ-plot above we could conclude that since not all the points fall on the dotted line, thus the residuals are not normal, it also appears to be heavy tailed.

(c) From the Residual vs. Fitted plot we can see that for each vertical line of points representing a different treatment, the spread on the points does not appear to be equal. It presents a trumpet pattern, indicating that these 3 treatments do not have the same variance, the 3rd treatment has vastly larger spread than the others. So the assumption of constant variance is violated.

Q3

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

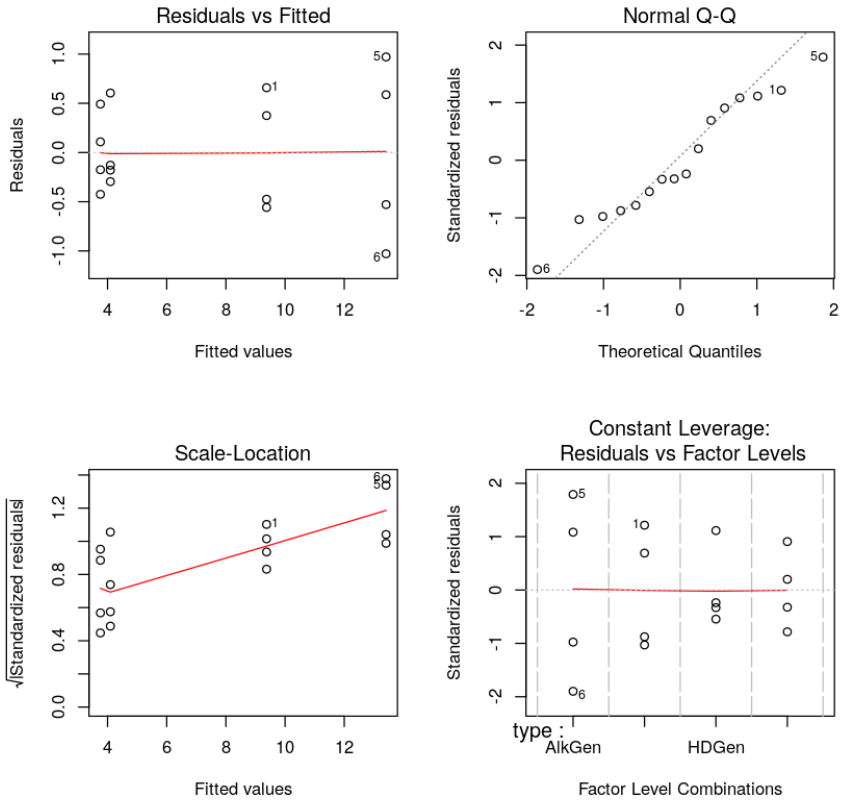
(a)

```
In [12]: batt$sqrt_life = sqrt(batt$life)
         aov.batt = aov(sqrt_life~type, data = batt)
         kable(anova(aov.batt), format='markdown')
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	3	255.835800	85.2785998	217.5284	0
Residuals	12	4.704412	0.3920344	NA	NA

(b)

```
In [13]: par(mfrow=c(2,2))
         plot(aov.batt)
```



From the QQ-plot above we could conclude that although most of the points fall on the dotted line, except for a few on the top right and lower left of the plot, thus the residuals are not normal.

(c) From the Residual vs. Fitted plot we can see that for each vertical line of points representing a different treatment, the spread on the points seems more evenly distributed around the horizontal baseline than the result before transformation, but the 3rd treatment still appears to be more spread-out than the other two, indicating that these 3 treatments do not have the same variance. So the assumption of constant variance is still violated.

Q4

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

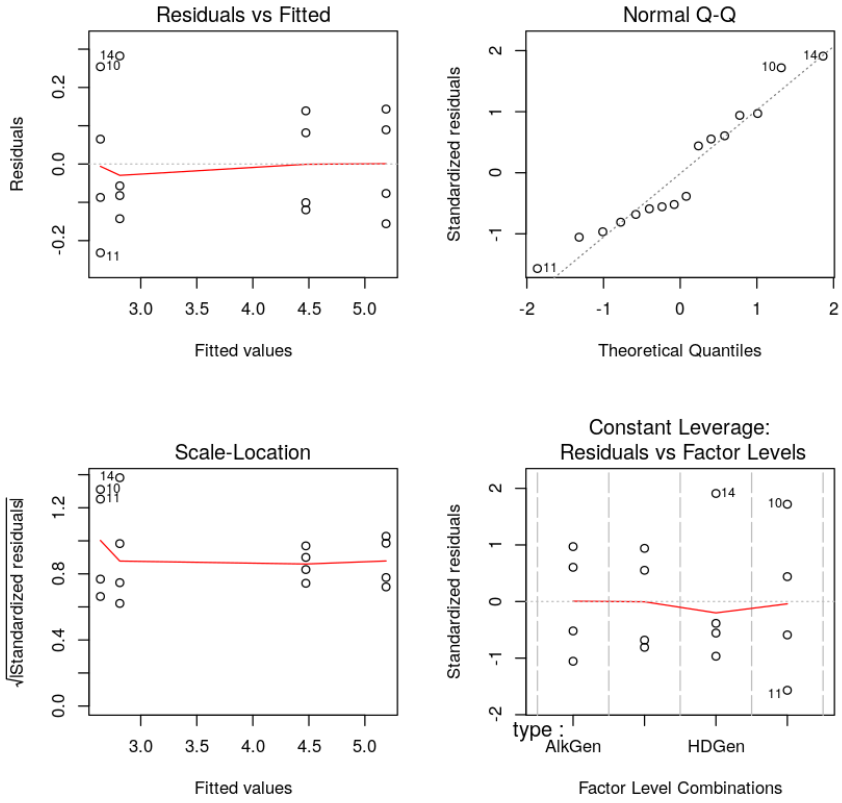
(a)

```
In [14]: batt$log_life = log(batt$life)
         aov.batt = aov(log_life~type, data = batt)
         kable(anova(aov.batt), format='markdown')
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	3	18.8002838	6.2667613	215.1637	0
Residuals	12	0.3495066	0.0291256	NA	NA

(b)

```
In [15]: par(mfrow=c(2,2))
         plot(aov.batt)
```



From the QQ-plot above we could conclude that basically all the points fall on the dotted line, thus the residuals are approximately normal.

(c) From the Residual vs. Fitted plot we can see that for each vertical line of points representing a different treatment, the spread on the points appears to be approximately equal, indicating that these 3 treatments have the same variance. So the assumption of constant variance is not violated.

Q5

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

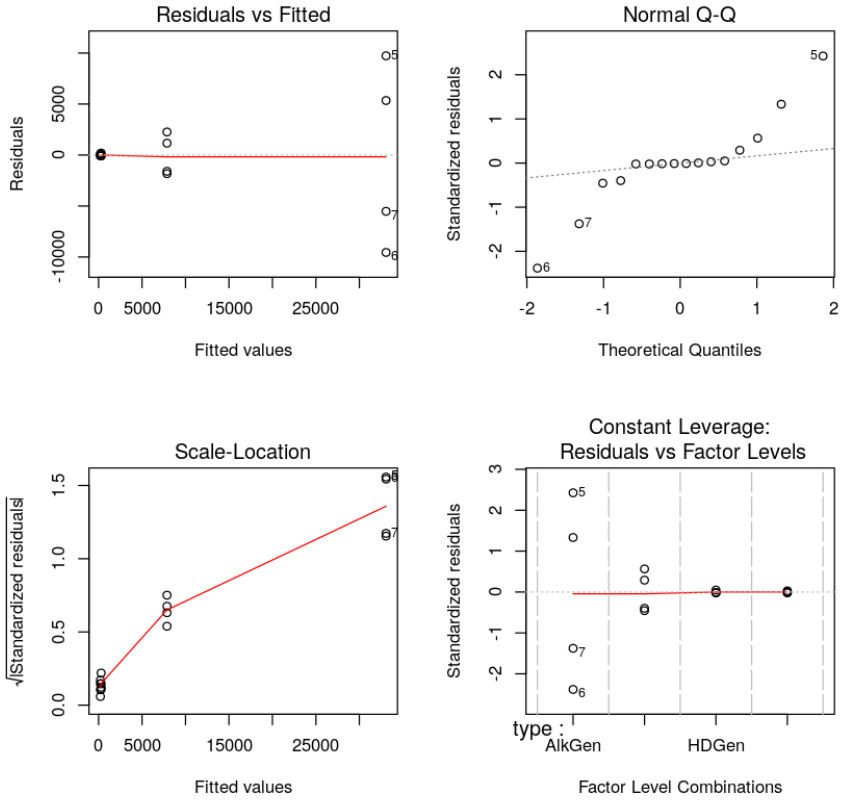
(a)

```
In [20]: batt$square_life = batt$life^2
         aov.batt = aov(square_life~type, data = batt)
         kable(anova(aov.batt), format='markdown')
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	3	2905095650	968365217	45.13527	8e-07
Residuals	12	257456832	21454736	NA	NA

(b)

```
In [21]: par(mfrow=c(2,2))
         plot(aov.batt)
```



From the QQ-plot above we could conclude that since not all the points fall on the dotted line, thus the residuals are not normal, it also appears to be VERY heavy tailed.

(c) From the Residual vs. Fitted plot we can see that for each vertical line of points representing a different treatment, the spread on the points does not appear to be equal. It presents a trumpet pattern, indicating that these 3 treatments do not have the same variance, the 3rd treatment has vastly larger spread than the others. So the assumption of constant variance is violated.

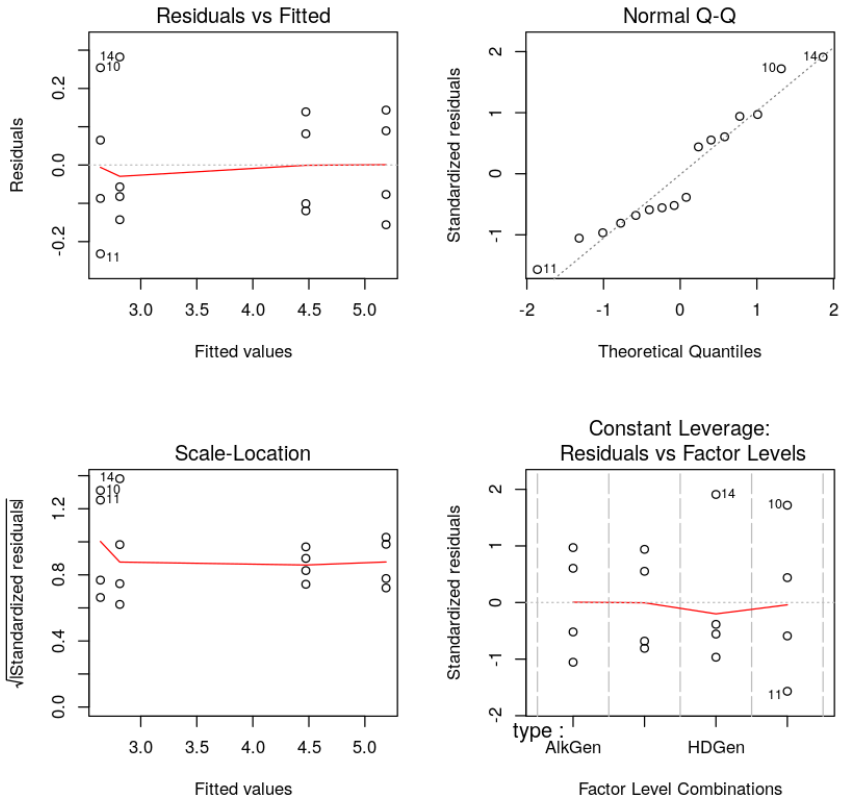
Q6

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

The log transformation appears to be the best for satisfying the normality and constant variance assumptions of residuals

```
In [24]: batt$log_life = log(batt$life)
         aov.batt = aov(log_life~type, data = batt)
         kable(anova(aov.batt), format='markdown')
         par(mfrow=c(2,2))
         plot(aov.batt)
```

```
|          | Df|     Sum Sq|   Mean Sq|  F value| Pr(>F)|
|:---------|--:|----------:|---------:|--------:|------:|
|type      |  3| 18.8002838| 6.2667613| 215.1637|      0|
|Residuals | 12|  0.3495066| 0.0291256|       NA|     NA|
```



- From the QQ-plot above we could conclude that basically all the points fall on the dotted line, thus the residuals are approximately normal
- From the Residual vs. Fitted plot we can see that for each vertical line of points representing a different treatment, the spread on the points seems more evenly distributed around the horizontal baseline than the result before transformation, indicating that these 3 treatments have the same variance. So the assumption of constant variance is not violated.

Q7

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.

Denote log battery life as l,

- Null hypothesis:

$$\text{For } Y_{it} = \mu + \tau_i + \epsilon_{it}, \text{where } i = AlkGen, AlkName, HDGen, HDName$$

$$H_0 : \tau_i = 0 \text{ for } i = AlkGen, AlkName, HDGen, HDName$$

```
In [48]: aov.batt = aov(log_life~type, data = batt)
lsm.batt=lsmeans(aov.batt, ~ type)
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	0.7157653	0.1206763	12	0.3574892	1.0740414	5.931281	3.480708e-04
AlkGen - HDGen	2.3758523	0.1206763	12	2.0175762	2.7341284	19.687807	1.011516e-09
AlkGen - HDName	2.5489199	0.1206763	12	2.1906438	2.9071960	21.121954	4.272094e-10
AlkName - HDGen	1.6600870	0.1206763	12	1.3018109	2.0183631	13.756526	5.448802e-08
AlkName - HDName	1.8331546	0.1206763	12	1.4748785	2.1914307	15.190672	1.761964e-08
HDGen - HDName	0.1730675	0.1206763	12	-0.1852085	0.5313436	1.434147	5.034253e-01

```
In [49]: AlkGen_AlkName_pv = 3.480708e-04
AlkGen_HDGen_pv = 1.011516e-09
AlkGen_HDName_pv = 4.272094e-10
AlkName_HDGen_pv = 5.448802e-08
AlkName_HDName_pv = 1.761964e-08
HDGen_HDName_pv = 5.034253e-01

AlkGen_AlkName_pv < 0.5
AlkGen_HDGen_pv < 0.5
AlkGen_HDName_pv < 0.5
AlkName_HDGen_pv < 0.5
AlkName_HDName_pv < 0.5
HDGen_HDName_pv < 0.5
```

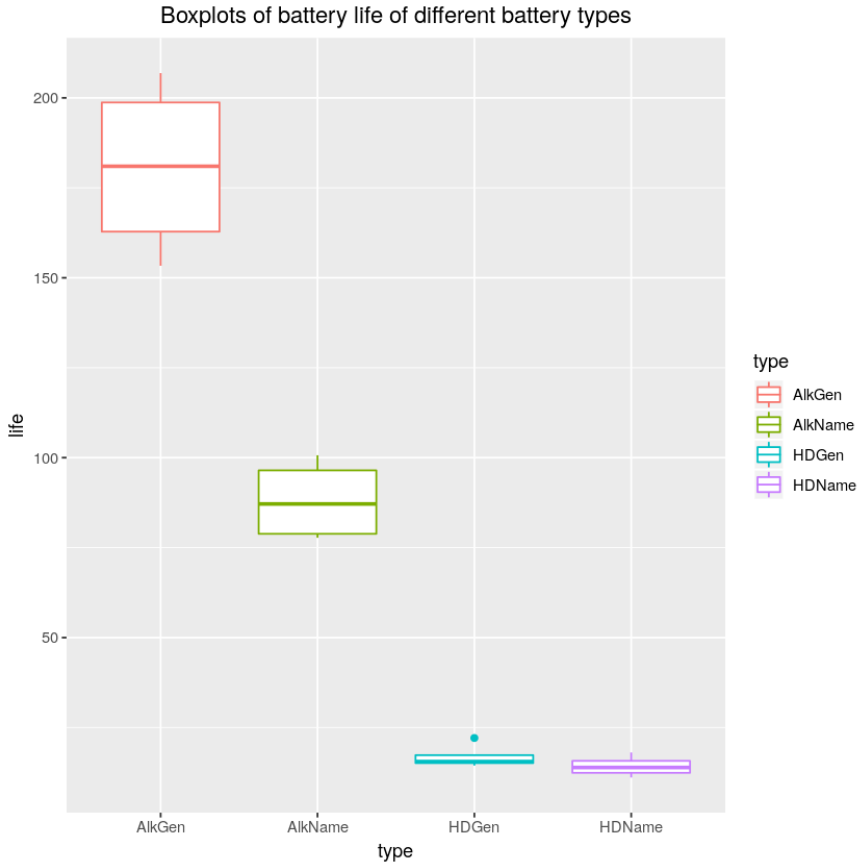
TRUE
TRUE
TRUE
TRUE
TRUE
FALSE

We can interpret the results of these tests with the following statements:

- 1. The battery life of AlkGen is significantly longer than the battery life of AlkName;
- 2. The battery life of AlkGen is significantly longer than the battery life of HDGen;
- 3. The battery life of AlkGen is significantly longer than the battery life of HDName;
- 4. The battery life of AlkName is significantly longer than the battery life of HDGen;
- 5. The battery life of AlkName is significantly longer than the battery life of HDName;
- 3. The battery life of HDGen is not significantly different from the battery life of HDName;

which we could validate with the boxplot:

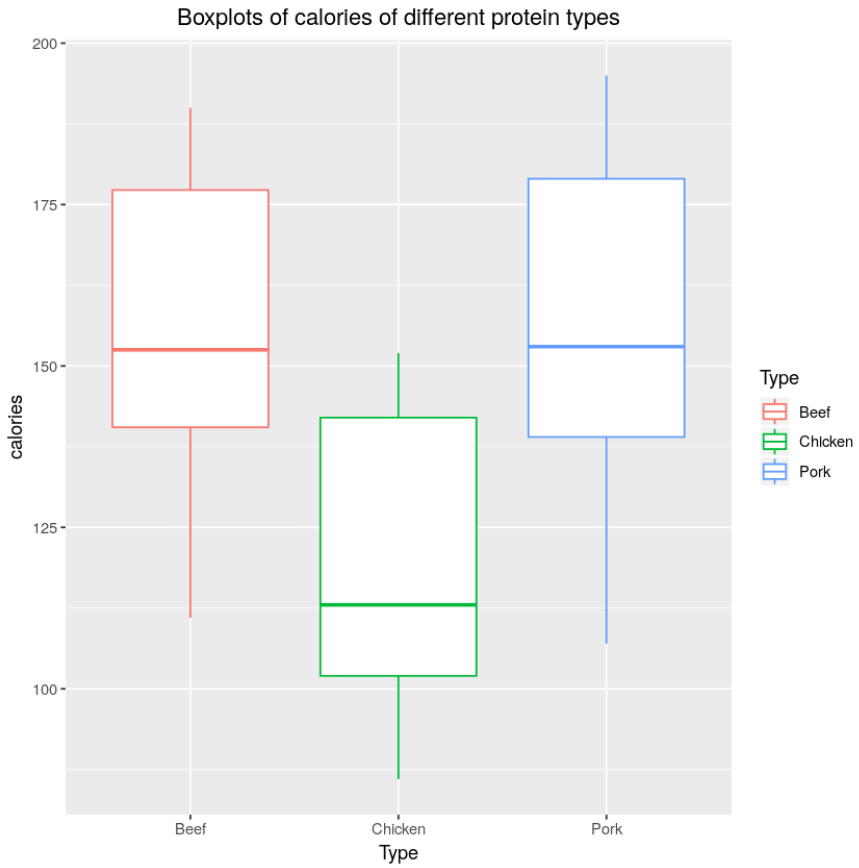
```
In [32]: ggplot(batt, aes(x=type, y=life, color=type)) +
                  geom_boxplot() +
                  ylab('life') +
                  ggtitle('Boxplots of battery life of different battery types') +
                  theme(plot.title = element_text(hjust = 0.5))
```



Q8

Hot Dogs A study was conducted to compare the calories and sodium in hot dogs made with different types of meat, 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.

```
In [35]: hotdog=read.table("hotdogs.txt",header=TRUE)
         ggplot(hotdog, aes(x=Type, y=Calories, color=Type)) +
         geom_boxplot() +
         ylab('calories') +
         ggtitle('Boxplots of calories of different protein types') +
         theme(plot.title = element_text(hjust = 0.5)) +
         ylim(min(hotdog$Calories)-0.05, max(hotdog$Calories)+0.05)
```



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.

Model and Null Hypothesis

$$Y_{it} = \mu + \tau_i + \epsilon_{it}, \text{where } i = beef, pork, chicken$$

$$\epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

and the number of replicates:

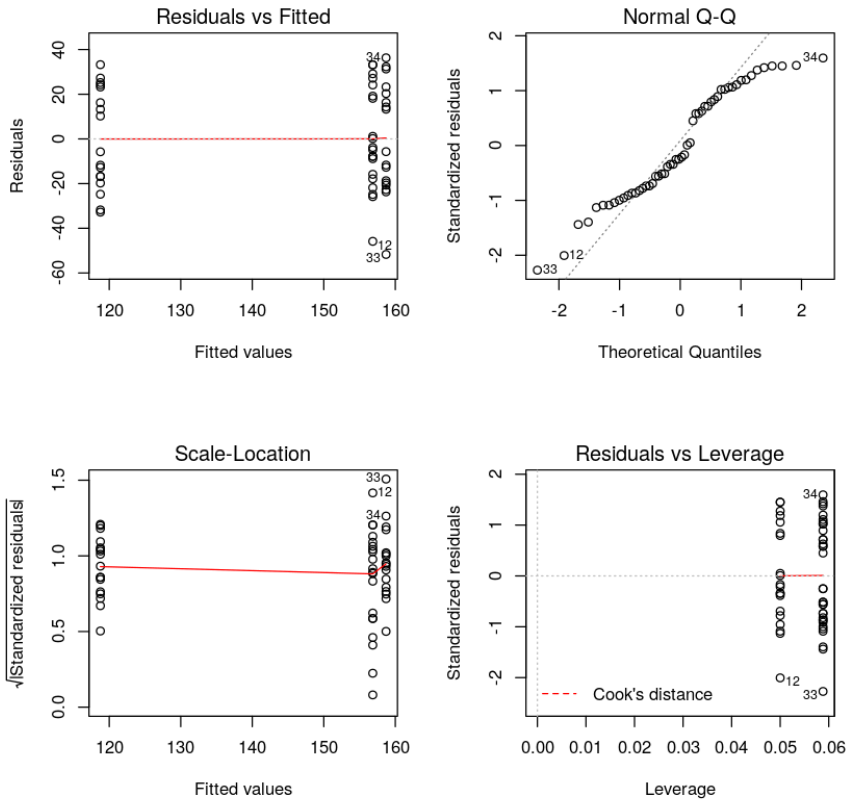
$$t_{beef} = 20, t_{pork} = 17, t_{chicken} = 17$$

we need to test:

$$H_0 : \tau_i = 0 \text{ for } i = beef, pork, chicken$$

Check Whether Transformation is Needed

```
In [46]: par(mfrow=c(2,2))
         aov.hotdog =aov(Calories~Type, data=hotdog)
          plot(aov.hotdog)
```



- From the QQ-plot above we could conclude that basically all the points fall on the dotted line, thus the residuals are approximately normal
- From the Residual vs. Fitted plot we can see that for each vertical line of points representing a different treatment, the spread on the points seems evenly distributed around the horizontal baseline, indicating that these 3 treatments have the same variance. So the assumption of constant variance is not violated.

Therefore no transformation needed.

ANOVA

```
In [43]:kable(anova(aov.hotdog), format='markdown')
        3.9e-06 < 0.5
```

```
|          | Df|   Sum Sq|  Mean Sq|  F value|  Pr(>F)|
|:---------|--:|--------:|--------:|--------:|-------:|
|Type      |  2| 17692.20| 8846.098| 16.07399| 3.9e-06|
|Residuals | 51| 28067.14|  550.336|       NA|      NA|
```

TRUE

so with an F-test from ANOVA table, we can see that since p-value = 3.862e-06$<$ 0.5, hence with statistical significance at the 0.05 level, we could conclude that there is difference in calories of different hotdogs.

Pairwise Difference

```
In [47]: lsm.hotdog=lsmeans(aov.hotdog, ~Type)
         summary(contrast(lsm.hotdog, method="pairwise",
                          adjust="tukey"),infer=c(T,T),
                          level=0.95, side="two-sided")

         Beef_Chicken_pv = 2.767694e-05
         Beef_Pork_pv = 9.688129e-01
         Chicken_Pork_pv = 2.390087e-05

         Beef_Chicken_pv < 0.5
         Beef_Pork_pv < 0.5
         Chicken_Pork_pv < 0.5
```

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Beef - Chicken	38.085294	7.738831	51	19.40391	56.76667	4.9213237	2.767694e-05
Beef - Pork	-1.855882	7.738831	51	-20.53726	16.82550	-0.2398143	9.688129e-01
Chicken - Pork	-39.941176	8.046454	51	-59.36515	-20.51720	-4.9638236	2.390087e-05

TRUE
FALSE
TRUE

We can interpret the results of these tests with the following statements:

- 1. The calories of beef is not significantly different from the calories of pork;
- 2. The calories of beef is significantly more than the calories of chicken;
- 3. The calories of chicken is significantly less than the calories of pork;