STAT 222 Spring 2021 HW6

Matthew Zhao

```
wtlift = read.table("http://www.stat.uchicago.edu/~yibi/s222/weight.lifting.txt", h=T)
```

Q1 — 7 points

```
library(mosaic)
mean(rate ~ 1, data=wtlift)
##       1
## 31.1364
mean(rate ~ A, data=wtlift)
##       1       2
## 32.2727 30.0000
mean(rate ~ B, data=wtlift)
##       1       2       3
## 26.5909 32.2273 34.5909
mean(rate ~ B+A, data=wtlift)
##     1.1     2.1     3.1     1.2     2.2     3.2
## 27.9091 33.0000 35.9091 25.2727 31.4545 33.2727
```

Using the code above, we estimate $hat\mu = \bar{y}_{...} = 31.1364$

$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}$
$\hat{\alpha}_1 = 32.2727 - 31.1364 = 1.1363$
$\hat{\alpha}_2 = 30.0000 - 31.1364 = -1.1364$

$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}$
$\hat{\beta}_1 = 26.5909 - 31.1364 = -4.5455$
$\hat{\beta}_2 = 32.2273 - 31.1364 = 1.0909$
$\hat{\beta}_3 = 34.5909 - 31.1364 = 3.4545$

$\hat{\alpha\beta}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$
$\hat{\alpha\beta}_{11} = 27.9091 - 32.2727 - 26.5909 + 31.1364 = 0.1819$
$\hat{\alpha\beta}_{12} = 33.0000 - 32.2727 - 32.2273 + 31.1364 = -0.3636$
$\hat{\alpha\beta}_{13} = 35.9091 - 32.2727 - 34.5909 + 31.1364 = 0.1819$
$\hat{\alpha\beta}_{21} = 25.2727 - 30.0000 - 26.5909 + 31.1364 = -0.1818$
$\hat{\alpha\beta}_{22} = 31.4545 - 30.0000 - 32.2273 + 31.1364 = 0.3636$

$\hat{\alpha\beta}_{23} = 33.2727 - 30.0000 - 34.5909 + 31.1364 = -0.1818$

Remark (not required):

```
wtlift$A = as.factor(wtlift$A)
wtlift$B = as.factor(wtlift$B)
contrasts(wtlift$A) = contr.sum(2)
contrasts(wtlift$B) = contr.sum(3)
lm1 = lm(rate ~ A * B, data=wtlift)
lm1$coef
## (Intercept)          A1          B1          B2       A1:B1       A1:B2
##   31.136364    1.136364   -4.545455    1.090909    0.181818   -0.363636
```

Q2 — 8 points

```
SST = var(wtlift$rate)*(66-1)
```

$$\text{SS}_\text{A} = bn \sum_{i=1}^{a} \hat{\alpha}_i^2$$
$$\text{SS}_\text{B} = an \sum_{j=1}^{b} \hat{\beta}_j^2$$
$$\text{SS}_\text{AB} = n \sum_{i=1}^{a} \sum_{j=1}^{b} \hat{\alpha\beta}_{ij}^2$$

```
wtlift %>% count(A,B)
##   A B  n
## 1 1 1 11
## 2 1 2 11
## 3 1 3 11
## 4 2 1 11
## 5 2 2 11
## 6 2 3 11
SSA = length(unique(wtlift$B)) * 11 * ((1.1363)^2 + (-1.1364)^2)
SSB = length(unique(wtlift$A)) * 11 * ((-4.5455)^2 + (1.0909)^2 + (3.4545)^2)
SSAB = 11 * ((0.1819)^2 + (-0.3636)^2 + (0.1819)^2 +
               (-0.1818)^2 + (0.3636)^2 + (-0.1818)^2)
SST - SSA - SSB - SSAB
## [1] 130.909
```

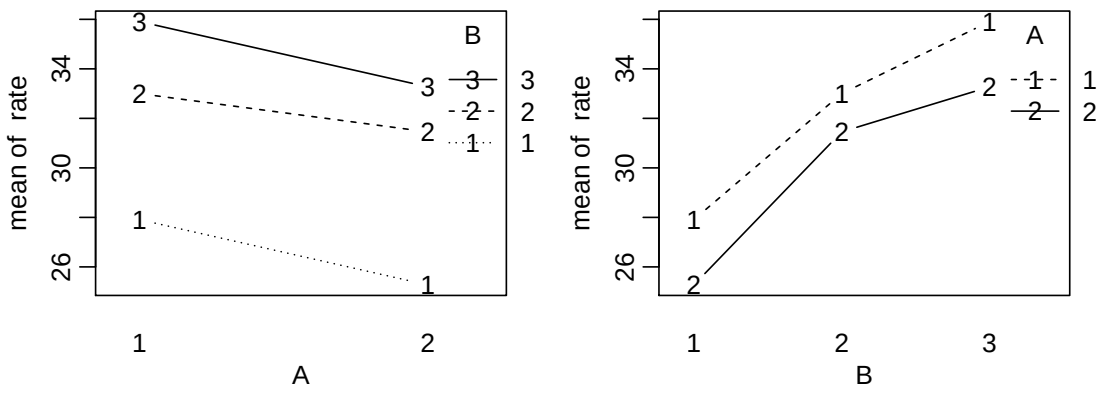
You may verify your calculation in R as follows:

```
lm1 = lm(rate ~ A * B, data=wtlift)
options(scipen=6, digits=8)  # increasing the number of digits in the output
anova(lm1)
```

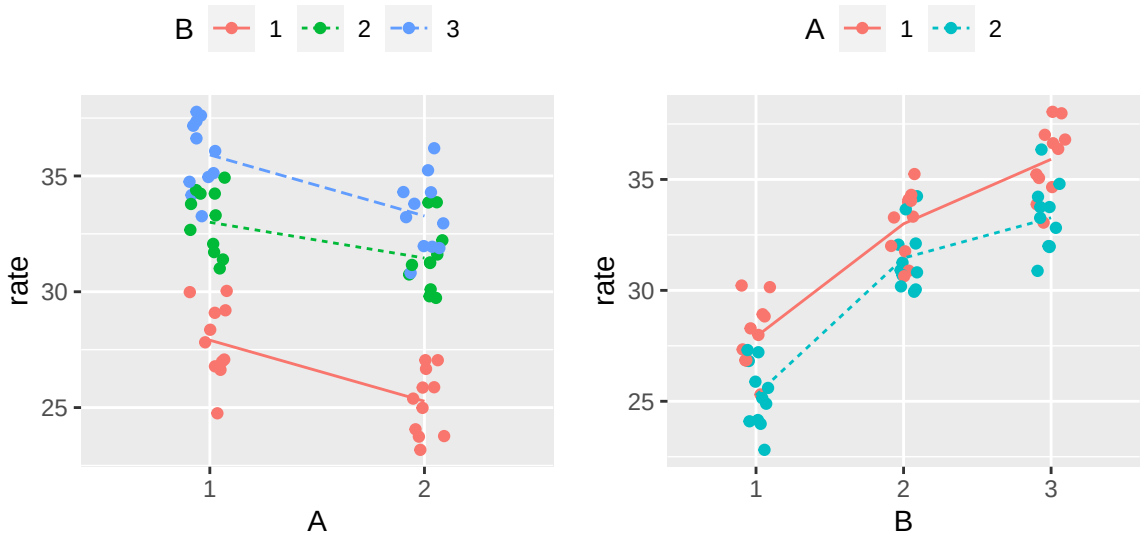
```
## Analysis of Variance Table
##
## Response: rate
##            Df  Sum Sq Mean Sq  F value         Pr(>F)
## A           1  85.227  85.227  39.0625 0.000000046796
## B           2 743.273 371.636 170.3333     < 2.22e-16
## A:B         2   4.364   2.182   1.0000        0.37393
## Residuals 60 130.909   2.182
options(scipen=6, digits=5)  # restoring the digit setting back to 5
```

Q3 — 3 points

```
par(mai=c(.6,.6,.1,.3),mgp=c(2,.6,0))
with(wtlift, interaction.plot(A, B, rate, type="b"))
with(wtlift, interaction.plot(B, A, rate, type="b"))
```



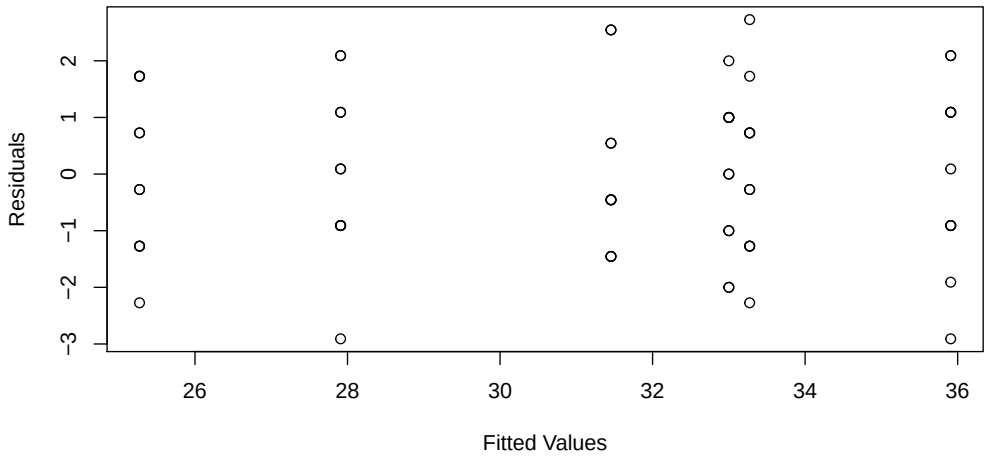

```
wtlift$A = as.factor(wtlift$A)
wtlift$B = as.factor(wtlift$B)
library(ggplot2)
ggplot(wtlift, aes(x=A,y=rate, color=B))+
  geom_point(position = position_jitter(width = .1))+
  stat_summary(fun="mean",geom="line",aes(group=B,linetype=B)) +
  theme(legend.position="top")
ggplot(wtlift, aes(x=B,y=rate,color=A))+
  geom_point(position = position_jitter(width = .1))+
  stat_summary(fun="mean",geom="line",aes(group=A,linetype=A)) +
  theme(legend.position="top")
```



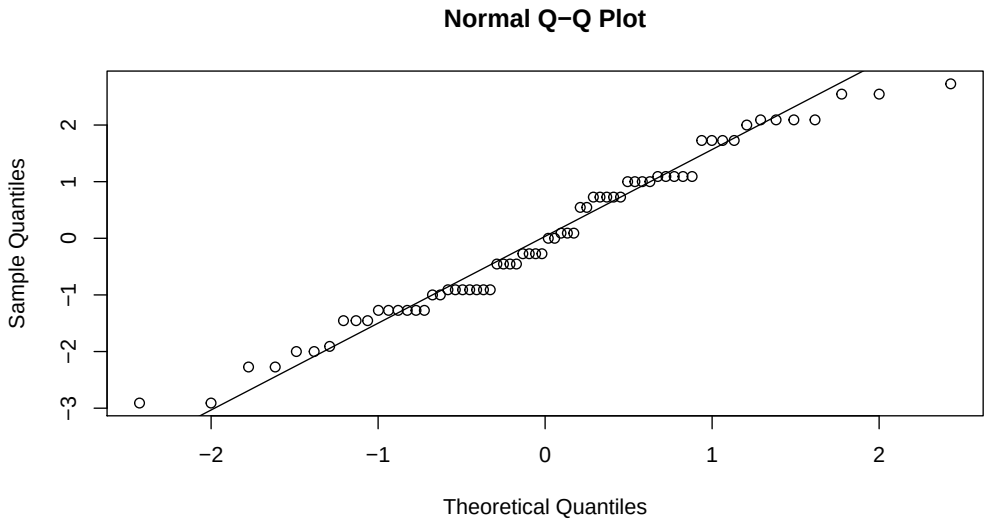
From the graphs, we can see that there is clearly a main effect for both A and B since varying the level of A or B while holding the other constant clearly results in different means. This makes sense since both main effects were statistically significant as seen in the ANOVA table with extremely low p values. Additionally, the graphs confirm that the interaction term is insignificant as seen in the ANOVA table with a high p value. The effects are roughly parallel in both the first and second graph, indicating that there is little interaction effect between A and B.

Q4 — 4 points

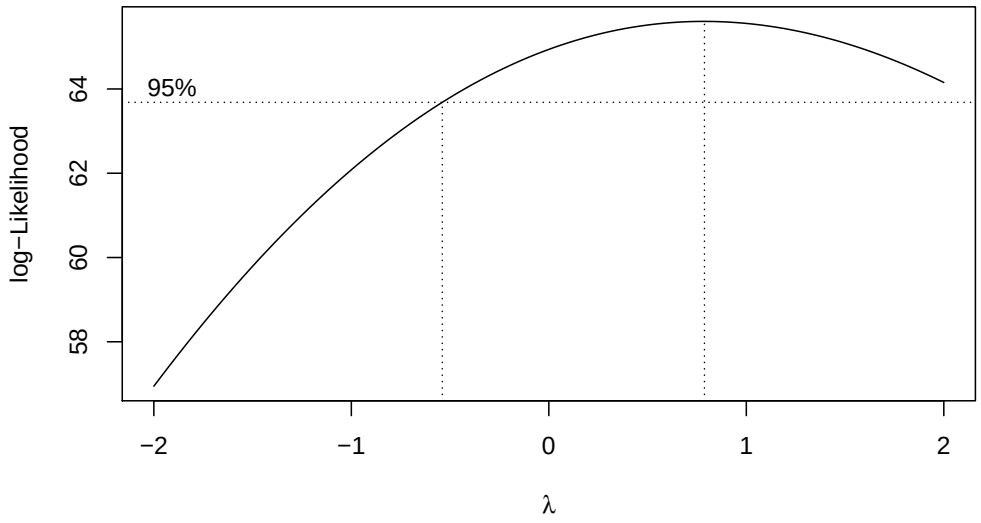
```
lm1 = lm(rate ~ A * B, data=wtlift)
plot(lm1$fitted.values, lm1$residuals, xlab = "Fitted Values", ylab = "Residuals")
```




```
qqnorm(lm1$residuals)
qqline(lm1$residuals)
```



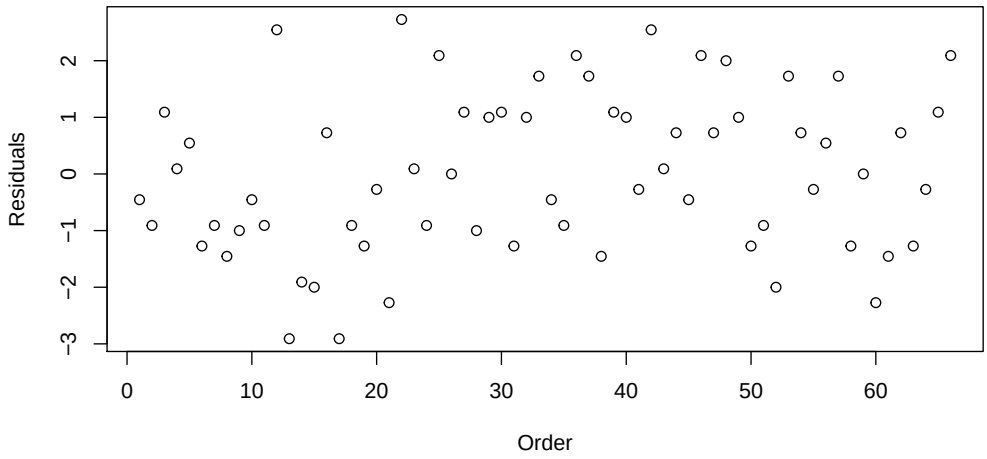

```
library(MASS)
boxcox(lm1)
```



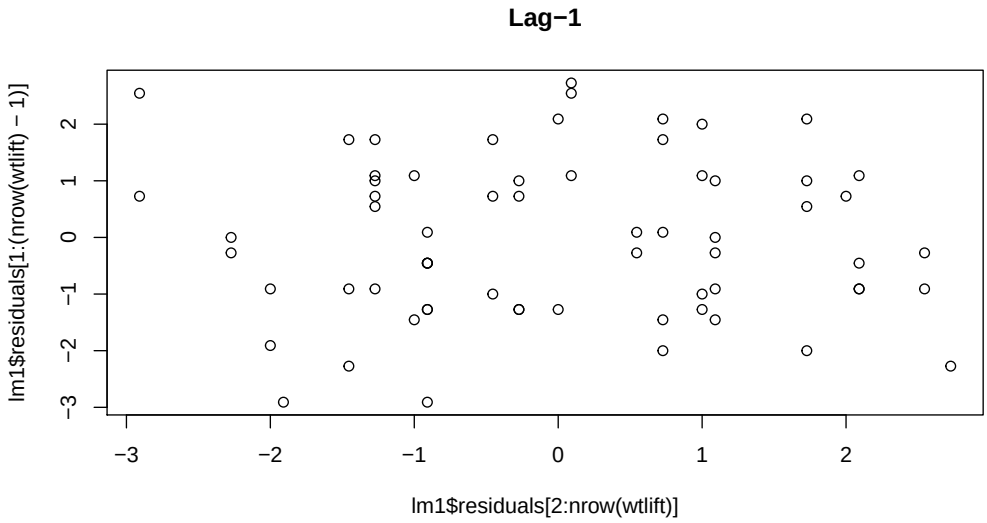
The plot of fitted values against the residuals reveals that the data could potentially be exhibiting non-constant variance. Additionally, the normal qq plot indicates that the data might not be normal since there is clustering in the center as well as points off the line at the edges. However, our boxcox plot shows that 1 does fall within our 95% confidence interval so we most likely do not need to perform a power or log transformation. I qualify this statement with the fact that the CI is extremely wide so it is possible that some transformation could be useful.

Q5 — 3 points

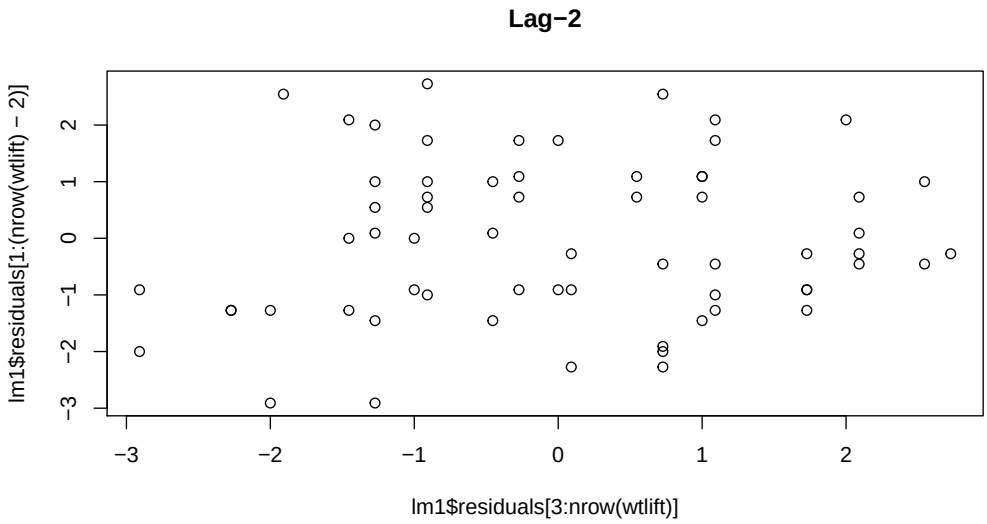
```
plot(wtlift$order, lm1$residuals, xlab = "Order", ylab = "Residuals")
```




```
plot(lm1$residuals[2:nrow(wtlift)], lm1$residuals[1:(nrow(wtlift)-1)], main = "Lag-1")
```




```
plot(lm1$residuals[3:nrow(wtlift)], lm1$residuals[1:(nrow(wtlift)-2)], main = "Lag-2")
```




```
acf(lm1$residuals, lag.max = 5,plot=F)
##
## Autocorrelations of series 'lm1$residuals', by lag
##
##      0      1      2      3      4      5
##  1.000 -0.019  0.093  0.121  0.154 -0.075
```

Based on the graphs generated, there does not appear to be any evidence of serial dependence among the residuals. The time plot along with the two lag plots show virtually no correlation, while acf does not contradict that, with autocorrelation values barely above and below zero.