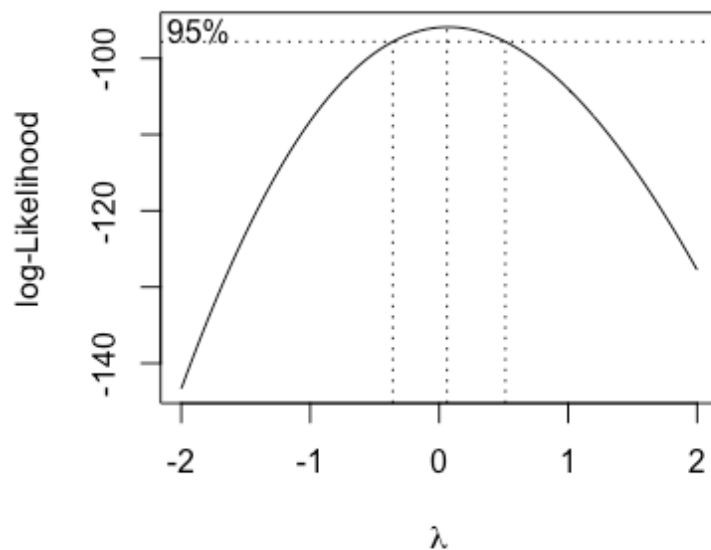# Homework 5

Name: Zixin Ouyang

1.

a.

```
library(alr4)
library(MASS)
> ais_mod<-lm(Wt~Ht,data=ais)
> boxcox(ais_mod)
```

The log transformation of Wt is suggested by the Box-Cox procedure. $\lambda=0$.

b.

```
> ais_mod1<-lm(log(Wt)~Ht, data=ais)
> sum(ais_mod1$residuals^2)
[1] 2.58568
> ais_mod2<-lm(log(Wt)~log(Ht), data=ais)
> sum(ais_mod2$residuals^2)
[1] 2.529865
```

The simple linear regression of log(Wt) on log(Ht) has smaller RSS.

c.

```
> ais_mod3<-lm(log(Wt)~log(Ht)+Sex+log(Ht)*Sex,data=ais)
> summary(ais_mod3)

Call:
lm(formula = log(Wt) ~ log(Ht) + Sex + log(Ht) * Sex, data = ais)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25413 -0.07049 -0.01276  0.05717  0.38787

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.4766     1.3532  -6.264 2.29e-09 ***
log(Ht)        2.4661     0.2591   9.517  < 2e-16 ***
Sex           -1.0036     1.8082  -0.555    0.580
log(Ht):Sex    0.1836     0.3480   0.528    0.598
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.111 on 198 degrees of freedom
Multiple R-squared:  0.6658,    Adjusted R-squared:  0.6607
F-statistic: 131.5 on 3 and 198 DF,  p-value: < 2.2e-16
```

The p-value of interaction term is 0.598, so it is insignificant.

d.
```
> plot(log(Wt)~log(Ht), type="n", data=ais)
> with(ais, points(log(Ht[Sex==0]), log(Wt[Sex==0]), pch=1, col="red", cex=0.8))
> with(ais, points(log(Ht[Sex==1]), log(Wt[Sex==1]), pch=2, col="blue", cex=0.8))
> beta00<--8.4766
> beta01<--8.4766-1.0036
> beta10<-2.4661
> beta11<-2.44661+0.1836
> abline(beta00, beta10, col="red",lwd=2, lty=1)
> abline(beta01, beta11, col="blue",lwd=2, lty=2)
> text(5.3, 4.7, "Sex=0", col="red", cex=1.0)
> text(5.1, 3.8, "Sex=1", col="blue", cex=1.0)
```
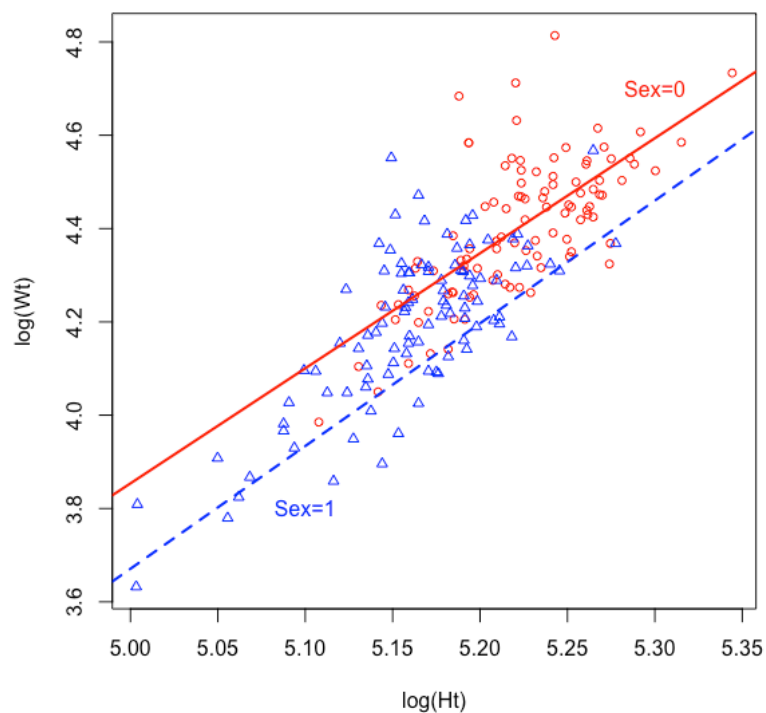
d.

```
> ais_mod4<-lm(log(Wt)~log(Ht)+Sex,data=ais)
> summary(ais_mod4)

Call:
lm(formula = log(Wt) ~ log(Ht) + Sex, data = ais)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25554 -0.07191 -0.01351  0.05763  0.38688

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.00810    0.90181  -9.989  < 2e-16 ***
log(Ht)      2.56790    0.17268  14.871  < 2e-16 ***
Sex         -0.04972    0.01881  -2.644  0.00884 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1108 on 199 degrees of freedom
Multiple R-squared:  0.6653,    Adjusted R-squared:  0.662
F-statistic: 197.8 on 2 and 199 DF,  p-value: < 2.2e-16
```

The p-value of Sex is less than 0.05, so it is significant.

e.
```
> plot(log(Wt)~log(Ht), type="n", data=ais)
> with(ais, points(log(Ht[Sex==0]), log(Wt[Sex==0]), pch=1, col="red", cex=0.8))
> with(ais, points(log(Ht[Sex==1]), log(Wt[Sex==1]), pch=2, col="blue", cex=0.8))
> beta00<--9.00810
> beta01<--9.00810-0.04972
> beta1<-2.56790
> abline(beta00, beta1, col="red",lwd=2, lty=1)
> abline(beta01, beta1, col="blue",lwd=2, lty=2)
> text(5.3, 4.7, "Sex=0", col="red", cex=1.0)
> text(5.1, 3.8, "Sex=1", col="blue", cex=1.0)
```
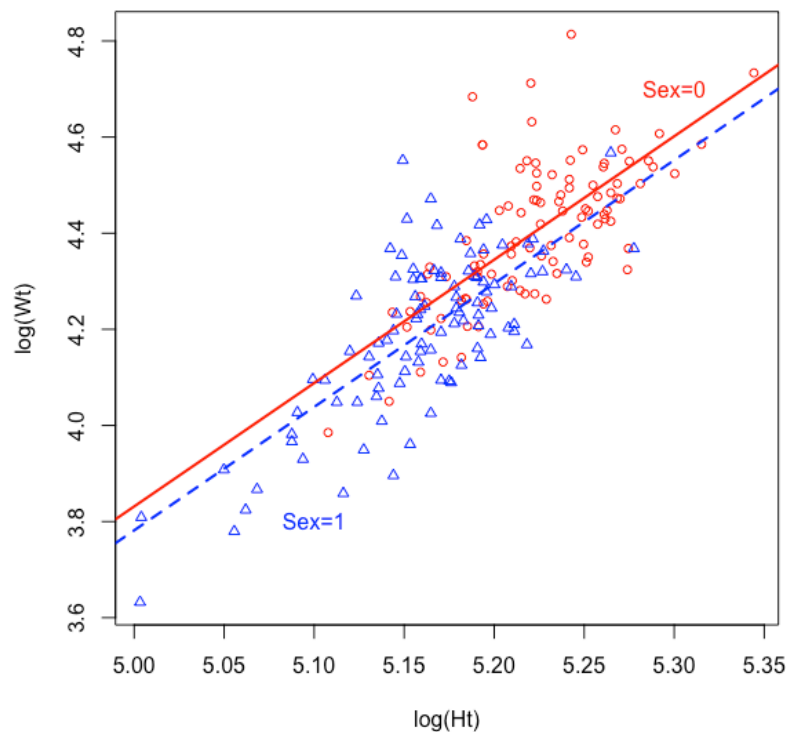

2.

a.
```
> table(turk0$A)
```

```
   0 0.04  0.1 0.16 0.28 0.44
  10    5    5    5    5    5
```
The design is not balanced because the first group has 10 experimental units, while other groups have 5 experimental units.

b.

```
> turk0_mod=lm(Gain~factor(A), data = turk0)
> summary(turk0_mod)

Call:
lm(formula = Gain ~ factor(A), data = turk0)

Residuals:
   Min    1Q Median    3Q    Max
 -33.4  -12.2   -0.6   13.6   38.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     623.00       5.82 107.041  < 2e-16 ***
factor(A)0.04    45.40      10.08   4.504 0.000101 ***
factor(A)0.1     92.60      10.08   9.186 4.37e-10 ***
factor(A)0.16   109.00      10.08  10.813 1.09e-11 ***
factor(A)0.28   171.00      10.08  16.963  < 2e-16 ***
factor(A)0.44   162.40      10.08  16.110 5.25e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.41 on 29 degrees of freedom
Multiple R-squared:  0.9386,     Adjusted R-squared:  0.928
F-statistic: 88.59 on 5 and 29 DF,  p-value: < 2.2e-16
```
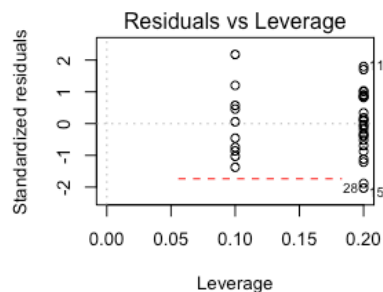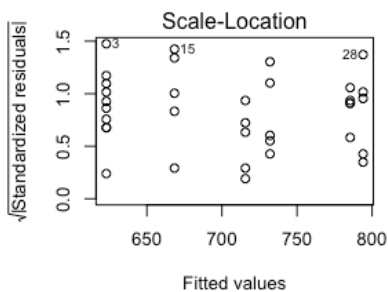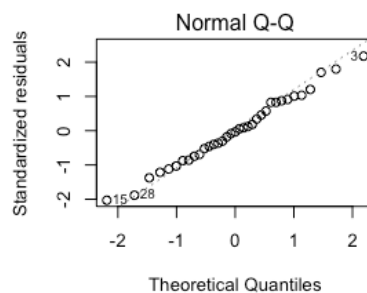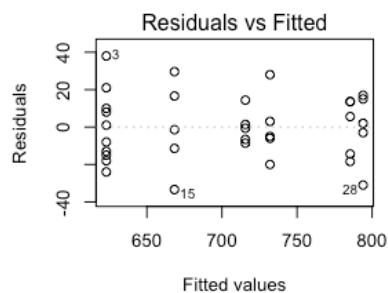
c.

```
> par(mfrow=c(2,2))
> plot(turk0_mod, add.smooth=FALSE)
```

The variance is constant and the normality assumption is valid.
d.
```
> anova(turk0_mod)
Analysis of Variance Table

Response: Gain
          Df Sum Sq Mean Sq F value    Pr(>F)
factor(A)  5 150041 30008.2  88.587 < 2.2e-16 ***
Residuals 29   9824   338.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e. The p-value in the anova table is smaller than 0.05, so there are differences among the mean weight gains of the groups.

f.
```
> TukeyHSD(aov(Gain~factor(A),data=turk0))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Gain ~ factor(A), data = turk0)

$`factor(A)`
             diff       lwr       upr     p adj
0.04-0       45.4  14.66875  76.13125 0.0012873
0.1-0        92.6  61.86875 123.33125 0.0000000
0.16-0      109.0  78.26875 139.73125 0.0000000
0.28-0      171.0 140.26875 201.73125 0.0000000
0.44-0      162.4 131.66875 193.13125 0.0000000
0.1-0.04     47.2  11.71461  82.68539 0.0042241
0.16-0.04    63.6  28.11461  99.08539 0.0000948
0.28-0.04   125.6  90.11461 161.08539 0.0000000
0.44-0.04   117.0  81.51461 152.48539 0.0000000
0.16-0.1     16.4 -19.08539  51.88539 0.7214957
0.28-0.1     78.4  42.91461 113.88539 0.0000030
0.44-0.1     69.8  34.31461 105.28539 0.0000222
0.28-0.16    62.0  26.51461  97.48539 0.0001380
0.44-0.16    53.4  17.91461  88.88539 0.0010275
0.44-0.28    -8.6 -44.08539  26.88539 0.9752675
```

g. The following pairs have significantly different means: (0, 0.04), (0, 0.1), (0, 0.16), (0, 0.28), (0, 0.44), (0.04, 0.1), (0.04, 0.16), (0.04, 0.28), (0.04, 0.44), (0.1, 0.28), (0.1, 0.44), (0.16, 0.28) and (0.16, 0.44).

3.

a.

```
> table(pine[,c(1,2)])
     trt
shape 1 2
    1 3 3
    2 3 3
    3 3 3
    4 3 3
```

There are eight treatment groups, and there are three experimental units in each group.

b.

```
> pine_mod<-lm(y~factor(shape)+factor(trt)+factor(shape)*factor(trt), data=pine)
> summary(pine_mod)

Call:
lm(formula = y ~ factor(shape) + factor(trt) + factor(shape) *
    factor(trt), data = pine)

Residuals:
   Min     1Q Median     3Q    Max
-9.333 -3.333 -1.167  4.250 11.667

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 11.333      3.877   2.924  0.00994 **
factor(shape)2              38.000      5.482   6.931 3.38e-06 ***
factor(shape)3              53.667      5.482   9.789 3.69e-08 ***
factor(shape)4              68.000      5.482  12.404 1.27e-09 ***
factor(trt)2                 4.667      5.482   0.851  0.40720
factor(shape)2:factor(trt)2 11.667      7.753   1.505  0.15187
factor(shape)3:factor(trt)2 11.333      7.753   1.462  0.16317
factor(shape)4:factor(trt)2 17.333      7.753   2.236  0.03998 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.714 on 16 degrees of freedom
Multiple R-squared:  0.9667,	Adjusted R-squared:  0.9522
F-statistic: 66.39 on 7 and 16 DF,  p-value: 1.241e-10
```
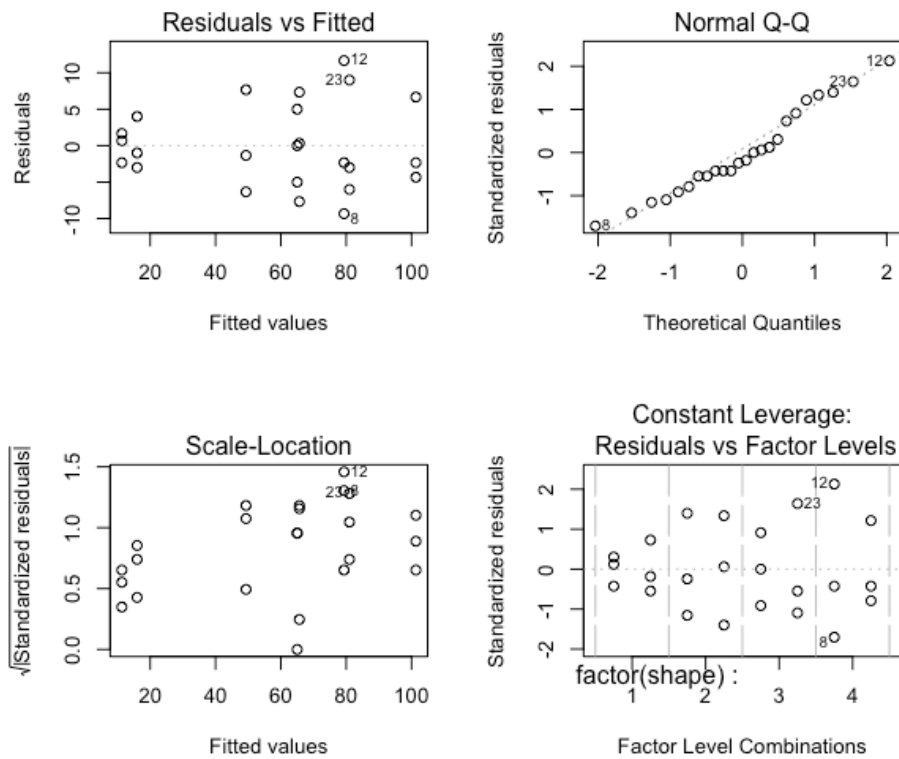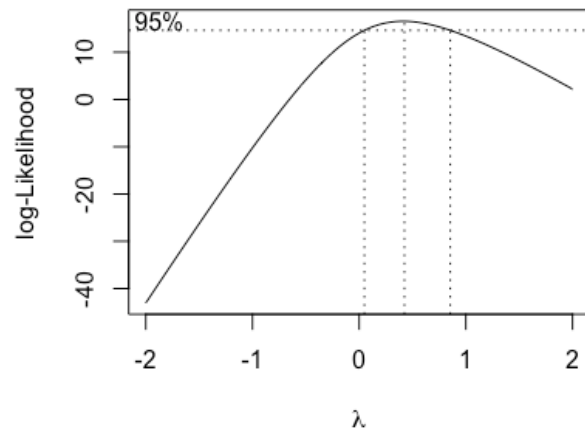
c.

```
> par(mfrow=c(2,2))
> plot(pine_mod,add.smooth=FALSE)
```

The plots show that the variance is not constant.

d.

```
> boxcox(pine_mod)
```

According to the plot above, we should choose $\lambda = \frac{1}{2}$. Transforming the response y into the square root of y is most appropriate.

e.

```
> pine_mod1<-lm(sqrt(y)~factor(shape)+factor(trt)+factor(shape)*factor(trt), data=pine)
> summary(pine_mod1)

Call:
lm(formula = sqrt(y) ~ factor(shape) + factor(trt) + factor(shape) *
    factor(trt), data = pine)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5271 -0.3174 -0.0971  0.3134  0.6457

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.3566     0.2491  13.474 3.77e-10 ***
factor(shape)2               3.6553     0.3523  10.376 1.64e-08 ***
factor(shape)3               4.7017     0.3523  13.346 4.35e-10 ***
factor(shape)4               5.5371     0.3523  15.717 3.79e-11 ***
factor(trt)2                 0.6270     0.3523   1.780   0.0941 .
factor(shape)2:factor(trt)2  0.4558     0.4982   0.915   0.3739
factor(shape)3:factor(trt)2  0.3077     0.4982   0.618   0.5456
factor(shape)4:factor(trt)2  0.5430     0.4982   1.090   0.2919
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4315 on 16 degrees of freedom
Multiple R-squared:  0.9763,    Adjusted R-squared:  0.9659
F-statistic: 94.11 on 7 and 16 DF,  p-value: 8.412e-12
```
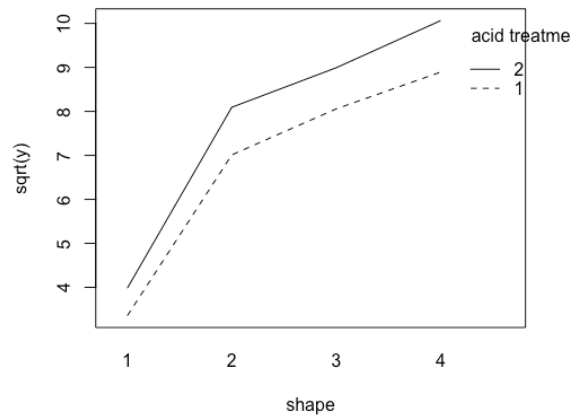
f.

```
> interaction.plot(pine$shape,pine$trt,sqrt(pine$y),trace.label='acid treatment',
+ xlab='shape',ylab='sqrt(y)')
```


g.
```
> anova(pine_mod1)
Analysis of Variance Table

Response: sqrt(y)
                          Df  Sum Sq Mean Sq  F value     Pr(>F)
factor(shape)              3 116.935  38.978 209.3714 4.786e-13 ***
factor(trt)                1   5.456   5.456  29.3088 5.739e-05 ***
factor(shape):factor(trt)  3   0.256   0.085   0.4581    0.7154
Residuals                 16   2.979   0.186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

h. The p-value for the interaction term is 0.7154, so we would conclude that the interaction term is not significant.

i. Because the interaction term is not significant and the experiment is balanced, it is appropriate to test for the main effects. According to the anova table above, shape and trt are significant with p-value less than 0.05.