Homework5

Huang Fang #913439658, Wenyu Li #913462971

February 19, 2016

```r
options(digits = 5)
```

Q26.4
a

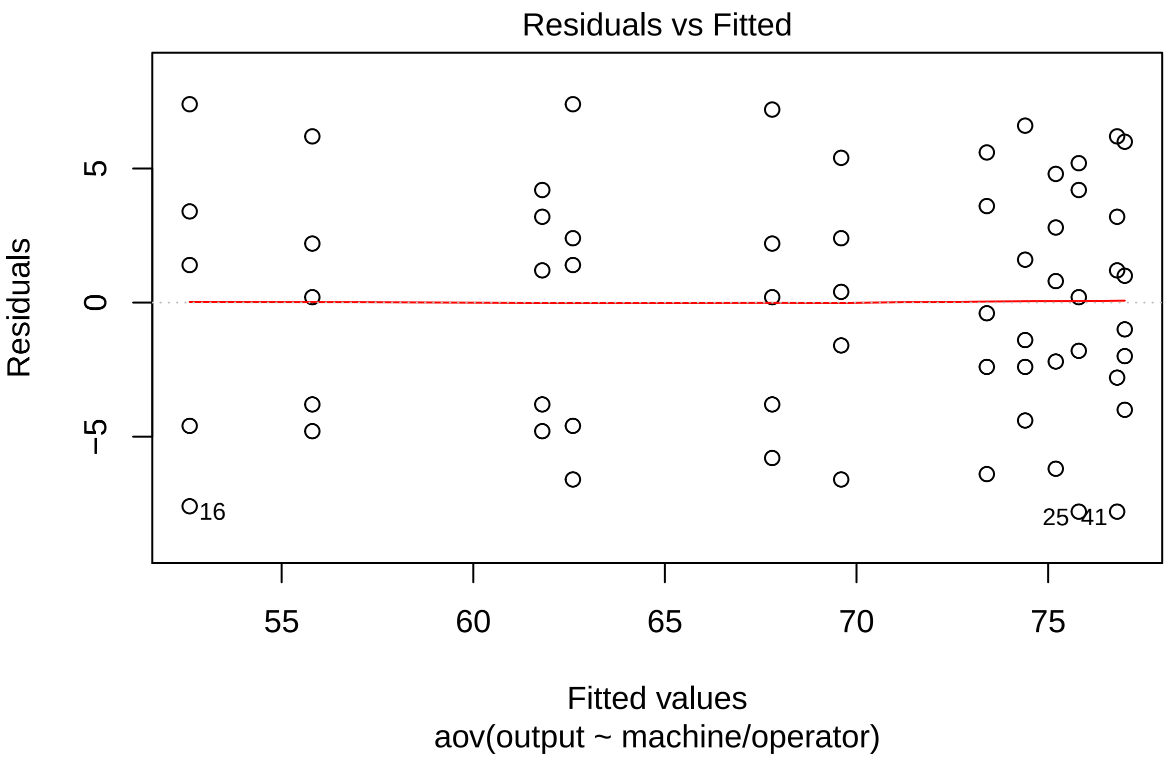
```r
machine = rep(c(1, 2, 3), each = 20)
operator = rep(c(1, 2, 3, 4), 3, each = 5)
output = c(65, 58, 63, 57, 66, 68, 62, 75, 64, 70, 56, 65, 58, 70, 64, 45, 56, 54, 48,
           60, 74, 81, 76, 80, 68, 69, 76, 80, 78, 73, 52, 56, 62, 58, 51, 73, 78, 83,
           75, 76, 69, 83, 74, 78, 80, 63, 70, 72, 68, 75, 81, 72, 73, 76, 70, 67, 79,
           73, 77, 71)
data = as.data.frame(cbind(machine, operator, output))
data$machine = as.factor(data$machine)
data$operator = as.factor(data$operator)
aov_out = aov(output ~ machine/operator, data = data)
```

The nested design model:
$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$
i = 1, 2, 3. j = 1, 2, 3, 4. k = 1, 2, 3, 4, 5.

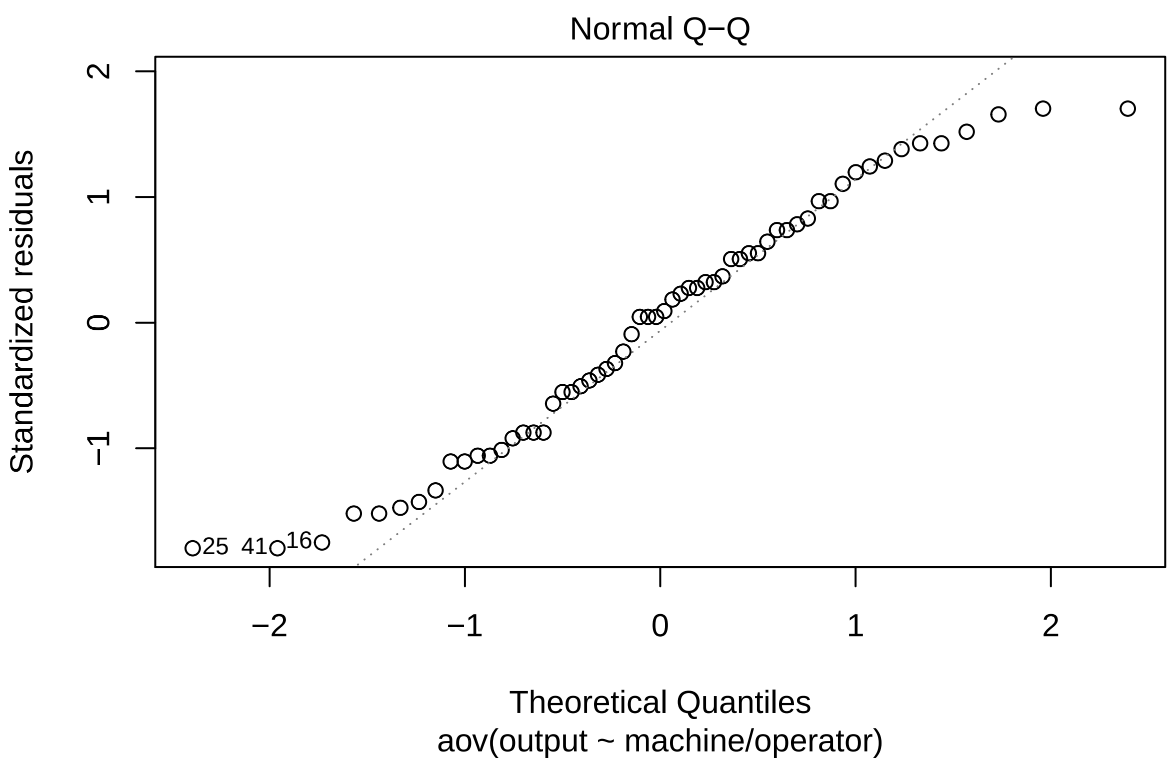
```r
aov_out$residuals
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##  3.2 -3.8  1.2 -4.8  4.2  0.2 -5.8  7.2 -3.8  2.2 -6.6  2.4 -4.6  7.4  1.4
##   16   17   18   19   20   21   22   23   24   25   26   27   28   29   30
## -7.6  3.4  1.4 -4.6  7.4 -1.8  5.2  0.2  4.2 -7.8 -6.2  0.8  4.8  2.8 -2.2
##   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45
## -3.8  0.2  6.2  2.2 -4.8 -4.0  1.0  6.0 -2.0 -1.0 -7.8  6.2 -2.8  1.2  3.2
##   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60
## -6.6  0.4  2.4 -1.6  5.4  6.6 -2.4 -1.4  1.6 -4.4 -6.4  5.6 -0.4  3.6 -2.4
```

```r
#par(mfrow = c(2, 1))
plot(aov_out, which = 1)
```



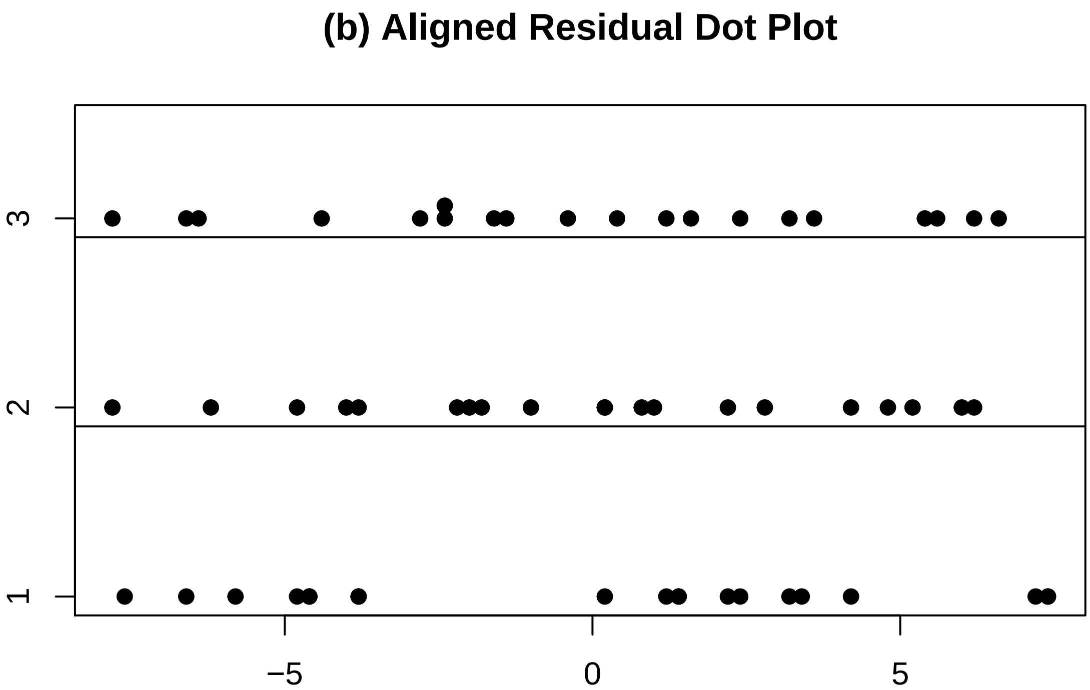

```r
plot(aov_out, which = 2)
```



From the plots, we find that there is no obvious nonlinear pattern in the Residual vs. Fitted plot, which means the nested design model seems to be approperiate. In the QQ plot, we can find that the distribution of residuals is approximately normal, although it seems a little bit light-tailed, the normality assumption holds.

b.

```
stripchart(split(resid(aov_out), data$machine), method = "stack",  pch = 19)
abline(h = seq(2, 4)-0.1)
title("(b) Aligned Residual Dot Plot")
```



From the aligned residual dot plots by machine, we find that the distribution of residuals is not affected by the value of machine, so the assumption of constancy of the error variance.

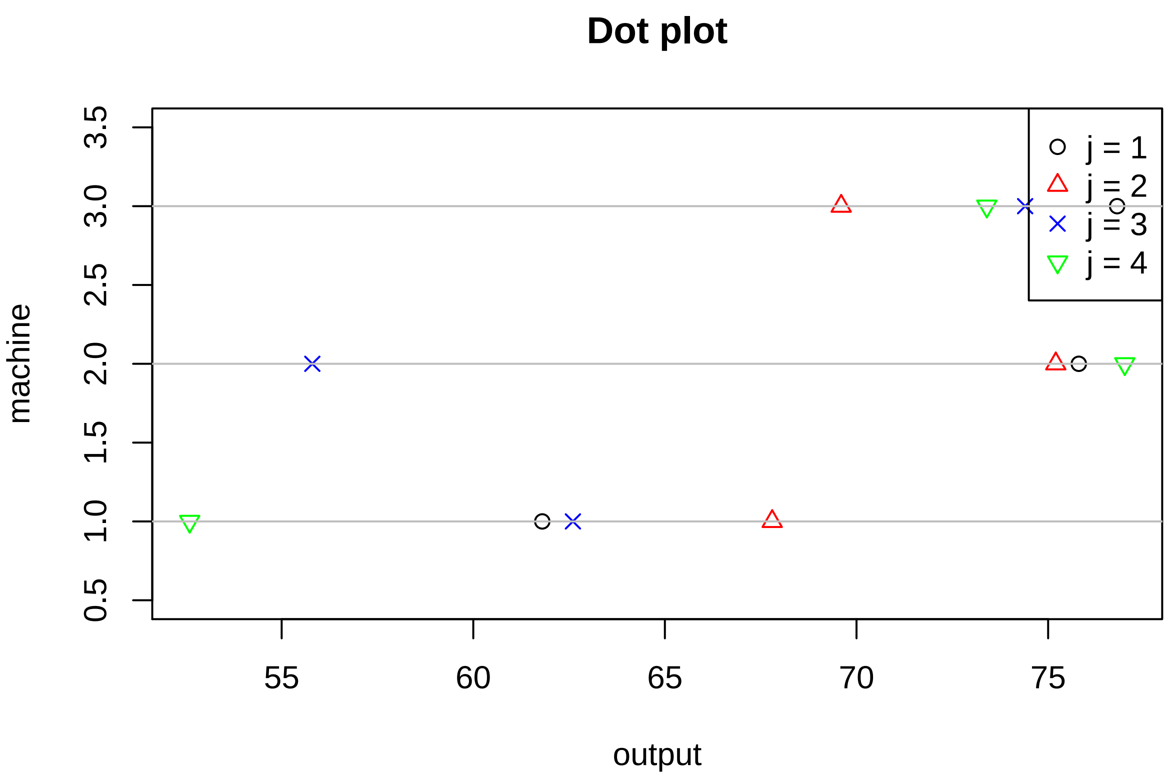
Q26.5

a.

No, the operator effect can't be distinguished from the effects of shifts.

b.

```
library(lattice)
par(mfrow = c(1,1))
result = with(data, tapply(output, list(machine, operator), mean))
plot(result[,1], c(1:3), xlab = "output", ylab = "machine", xlim = range(result),
     ylim = c(0.5, 3.5), pch = 1, main = "Dot plot")
points(result[,2], c(1:3), xlab = "output", pch = 2, col = "red")
points(result[,3], c(1:3), xlab = "output", pch = 4, col = "blue")
points(result[,4], c(1:3), pch = 6, col = "green" )
legend("topright", c("j = 1", "j = 2", "j = 3", "j = 4"), pch = c(1, 2, 4, 6),
       col = c("black", "red", "blue", "green"))
abline(h = seq(1, 3), col = "grey")
```




```
#dotplot(result, xlab = "output", ylab = "operator", main = "Dot plot")
```

From the dotplot, we find that the main effect of machine seems to be siginificant, and the main effect of operator seems to be significant under different levels of machine.

c.

```
summary(aov_out)
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## machine            2   1696     848    35.9 2.9e-10 ***
## machine:operator   9   2272     252    10.7 7.0e-09 ***
## Residuals         48   1133      24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d.

$H_0 : \alpha_i = 0 \; vs. \; H_a : exist \; \alpha_i \neq 0$
$F^* = MSA/MSE = 848/24 = 35.9$
If $F^* > F(0.99, 2, 48)$, we reject H_0, otherwise we conclude H_0.
$F^* = 35.9 > 5.08 = F(0.99, 2, 48)$
We reject H_0, there exist $\alpha_i \neq 0$.
The P-value is 2.9e-10.

e.

$H_0 : \beta_{j(i)} = 0 \; vs. \; H_a : exist \; \beta_{j(i)} \neq 0$
$F^* = MSB(A)/MSE = 252/24 = 10.7$
If $F^* > F(0.99, 9, 48)$, we reject H_0, otherwise we conclude H_0.
$F^* = 10.7 > 2.8 = F(0.99, 9, 48)$

We reject H_0, there exist $\beta_{j(i)} \neq 0$.
The P-value is 7.0e-09.
There exist $\beta_{j(i)} \neq 0$, but for i = 3, if there exist $\beta_{j(3)} \neq 0$, maybe the mean output for the four operators assigned to machine 3 differ, maybe not.

f.

```
n = 5
#all_means = with(data, tapply(output, list(machine), mean))
t = with(data, tapply(output, list(machine, operator), mean))
f1 = function(x){
  return(sum((x - mean(x))^2))
}
result = n*apply(t, 1, f1)
names(result) = c("SSB(A1)", "SSB(A2)", "SSB(A3)")
result
```

```
## SSB(A1) SSB(A2) SSB(A3)
##  599.20 1538.55  134.55
```

So, $SSB(A_1) = 599.20, SSB(A_2) = 1538.55, SSB(A_3) = 134.55$.
For i = 1:
$H_0 : \beta_{j(1)} = 0 \; vs. \; exist \; \beta_{j(1)} \neq 0$
$F^* = MSB(A_1)/MSE = 559.20/24 = 23.3$
If $F^* > F(0.99, 3, 48)$, we reject H_0, otherwise we conclude H_0.
$F^* = 23.3 > 4.218 = F(0.99, 3, 48)$
We reject H_0, there exist $\beta_{j(1)} \neq 0$.

For i = 2:
$H_0 : \beta_{j(2)} = 0 \; vs. \; exist \; \beta_{j(2)} \neq 0$
$F^* = MSB(A_2)/MSE = 1538.55/24 = 64.106$
If $F^* > F(0.99, 3, 48)$, we reject H_0, otherwise we conclude H_0.
$F^* = 64.106 > 4.218 = F(0.99, 3, 48)$
We reject H_0, there exist $\beta_{j(2)} \neq 0$.

For i = 3:
$H_0 : \beta_{j(3)} = 0 \; vs. \; exist \; \beta_{j(3)} \neq 0$
$F^* = MSB(A_3)/MSE = 134.55/24 = 5.6063$
If $F^* > F(0.99, 3, 48)$, we reject H_0, otherwise we conclude H_0.
$F^* = 5.6063 > 4.218 = F(0.99, 3, 48)$
We reject H_0, there exist $\beta_{j(3)} \neq 0$.

g.??
By Bonferroni inequality.
$\alpha = 1 - 0.99^5 = 0.04901$
Our set of conclusions:
There exist $\alpha_i \neq 0$.
There exist $\beta_{j(i)} \neq 0$.
There exist $\beta_{j(1)} \neq 0$.
There exist $\beta_{j(2)} \neq 0$.
There exist $\beta_{j(3)} \neq 0$.

Q26.6
a.

```
with(data, tapply(output, list(machine), mean))
```

```
##     1     2     3
## 61.20 70.95 73.55
```

T = qtukey(0.95, 3, 48)/sqrt(2) mse_root = 4.858
For $\mu_1 - \mu_2$, our 95% CI is $-9.75 \pm 2.4185 \times 4.858/\sqrt{10}$, which is (-13.465, -6.0347)
For $\mu_1 - \mu_3$, our 95% CI is $-12.35 \pm 2.4185 \times 4.858/\sqrt{10}$, which is (-16.065, -8.6347)
For $\mu_2 - \mu_3$, our 95% CI is $-2.6 \pm 2.4185 \times 4.858/\sqrt{10}$, which is (-6.3153, 1.1153)
μ_1 seems differ siginificantly from μ_2 and μ_3, and the difference between μ_2 and μ_2 is not significant.

b.

```
with(data, tapply(output, list(machine, operator), mean))[1, ]
```

```
##    1    2    3    4
## 61.8 67.8 62.6 52.6
```

There are total 6 pairs to compare. so B = qt(1 - 0.05/(2*6), 48) = 2.752, $s(\hat{L})$ = mse_root/sqrt(5/2) = 3.0725
For $\hat{L_1} = Y_{11\cdot} - Y_{12\cdot}$, its CI is $-6 \pm 2.752 \times 3.0725$, which is (-14.456, 2.4555)
For $\hat{L_2} = Y_{11\cdot} - Y_{13\cdot}$, its CI is $-0.8 \pm 2.752 \times 3.0725$, which is (-9.2555, 7.6555)
For $\hat{L_3} = Y_{11\cdot} - Y_{14\cdot}$, its CI is $9.2 \pm 2.752 \times 3.0725$, which is (0.7445, 17.655)
For $\hat{L_4} = Y_{12\cdot} - Y_{13\cdot}$, its CI is $5.2 \pm 2.752 \times 3.0725$, which is (-3.2555, 13.665)
For $\hat{L_5} = Y_{12\cdot} - Y_{14\cdot}$, its CI is $15.2 \pm 2.752 \times 3.0725$, which is (6.7445, 23.655)
For $\hat{L_6} = Y_{13\cdot} - Y_{14\cdot}$, its CI is $10 \pm 2.752 \times 3.0725$, which is (1.5445, 18.456)
$(\beta_{11}, \beta_{14}), (\beta_{12}, \beta_{14}) and (\beta_{13}, \beta_{14})$ these three pairs are significantly different, and we also find that both pairs include β_{14}, which means that β_{14} is different from other $\beta_{1i}, i \neq 4$

c.

```
with(data, tapply(output, list(machine, operator), mean))[1, ]
```

```
##    1    2    3    4
## 61.8 67.8 62.6 52.6
```

$L = \frac{\mu_{11}+\mu_{12}+\mu_{13}}{3} - \mu_{14}$
$\hat{L} = \frac{Y_{11\cdot}+Y_{12\cdot}+Y_{13\cdot}}{3} - Y_{14\cdot}$
$s(\hat{L}) = \sqrt{\frac{4}{3}\frac{MSE}{5}} = 2.5087$
$\hat{L} = (61.8 + 67.8 + 62.6)/3 - 52.6 = 11.467$
So the 95% CI is $11.467 \pm qt(0.995, 48) \times 2.5087$ which is (4.7382, 18.196).
The probability that (4.7382, 18.196) includes L is greater than 99%.

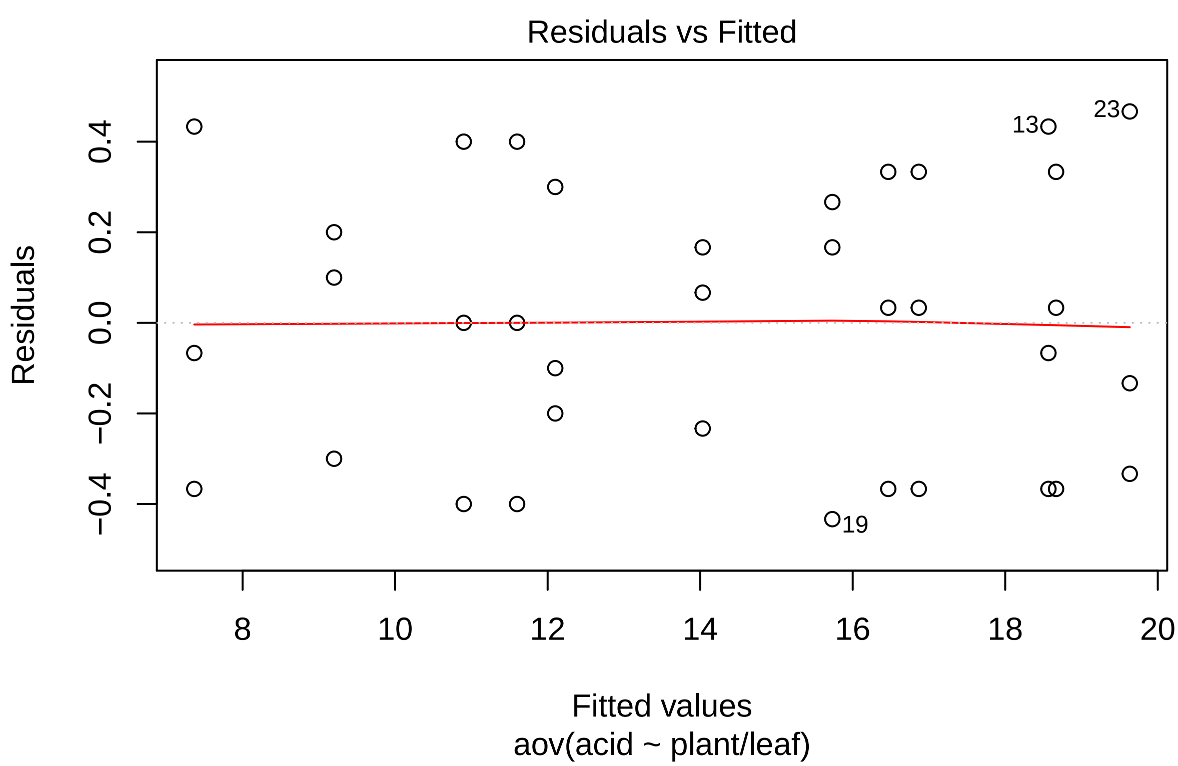
Q26.19

```
plant = rep(1:4, each = 9)
leaf = rep(1:3, 4, each = 3)
acid = c(11.2, 11.6, 12.0, 16.5, 16.8, 16.1, 18.3, 18.7, 19.0, 14.1, 13.8, 14.2,
         19.0, 18.5, 18.2, 11.9, 12.4, 12.0, 15.3, 15.9, 16.0, 19.5, 20.1, 19.3,
         16.5, 17.2, 16.9, 7.3, 7.8, 7.0, 8.9, 9.4, 9.3, 11.3, 10.9, 10.5)
data = as.data.frame(cbind(plant, leaf, acid))
```

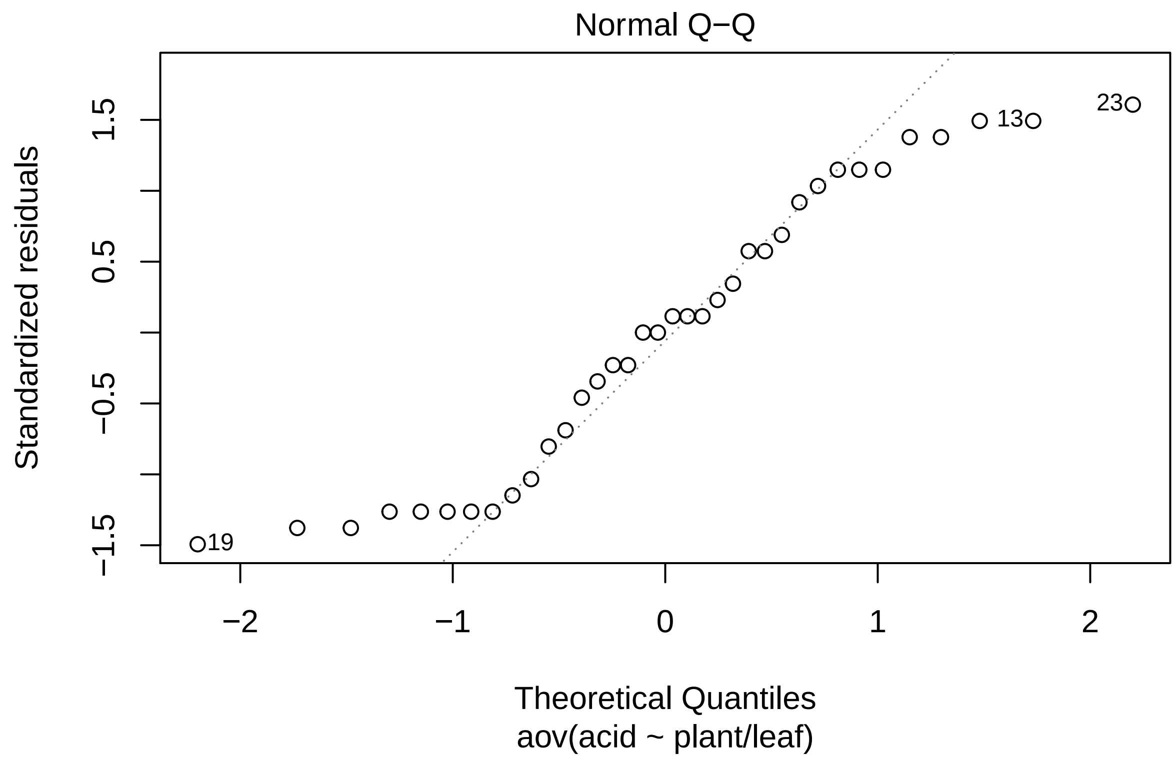
```r
data$plant = as.factor(data$plant)
data$leaf = as.factor(data$leaf)
aov_out = aov(acid ~ plant/leaf, data = data)
aov_out$residuals
```

```
##           1           2           3           4           5           6
## -4.0000e-01  7.3830e-15  4.0000e-01  3.3333e-02  3.3333e-01 -3.6667e-01
##           7           8           9          10          11          12
## -3.6667e-01  3.3333e-02  3.3333e-01  6.6667e-02 -2.3333e-01  1.6667e-01
##          13          14          15          16          17          18
##  4.3333e-01 -6.6667e-02 -3.6667e-01 -2.0000e-01  3.0000e-01 -1.0000e-01
##          19          20          21          22          23          24
## -4.3333e-01  1.6667e-01  2.6667e-01 -1.3333e-01  4.6667e-01 -3.3333e-01
##          25          26          27          28          29          30
## -3.6667e-01  3.3333e-01  3.3333e-02 -6.6667e-02  4.3333e-01 -3.6667e-01
##          31          32          33          34          35          36
## -3.0000e-01  2.0000e-01  1.0000e-01  4.0000e-01 -1.2490e-16 -4.0000e-01
```

```r
plot(aov_out, which = 1)
```




```r
plot(aov_out, which = 2)
```




```
#qq = qqnorm(aov_out$residuals)
#cov(sort(qq$x), sort(qq$y))
expvalue = rep(1, 36)
for(i in 1:36)
{
  expvalue[i] = qnorm((i - 0.375)/(36 + 0.25), 0, 1)
}
cor(sort(aov_out$residuals, decreasing = FALSE), expvalue)
```

```
## [1] 0.96853
```

From the plots, we find that there is no obvious nonlinear pattern in the Residual vs. Fitted plot, which means the nested design model seems to be approperiate. In the QQ plot, we can find that the distribution of residuals is approximately normal, although it seems a little bit light-tailed, the normality assumption holds.

Q26.20

a.

```
summary(aov_out)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## plant         3    343   114.4     905 <2e-16 ***
## plant:leaf    8    187    23.4     185 <2e-16 ***
## Residuals    24      3     0.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b.

$H_0 : \sigma_\alpha^2 = 0 \ vs. \ H_a : otherwise$

$F^* = MSA/MSB(A) = 4.8889$
If $F^* > F(0.95, 3, 8)$, we reject H_0, otherwise we conclude H_0
$F^* = 4.8889 > 4.0662 = F(0.95, 3, 8)$
So we reject H_0, $\sigma_\alpha^2 > 0$
P-value = 0.032324.

c.
$H_0 : \sigma_{B(A)}^2 = 0 \ vs. \ H_a : otherwise$
$F^* = MSB(A)/MSE = 185$
If $F^* > F(0.95, 8, 24)$, we reject H_0, otherwise we conclude H_0
$F^* = 185 > 2.3551 = F(0.95, 8, 24)$
So we reject H_0, $\sigma_{B(A)}^2 > 0$
P-value < <2e-16

d.
$\hat{\mu}_{...} = \bar{Y}_{...} = 14.261$
$s^2(\bar{Y}_{...}) = \frac{MSA}{nab} = 114.4/36 = 3.1778$
So the 95% CI is $14.261 \pm qt(0.975, df(MSA)) \times \sqrt{3.1778}$, which is (8.5878, 19.934)

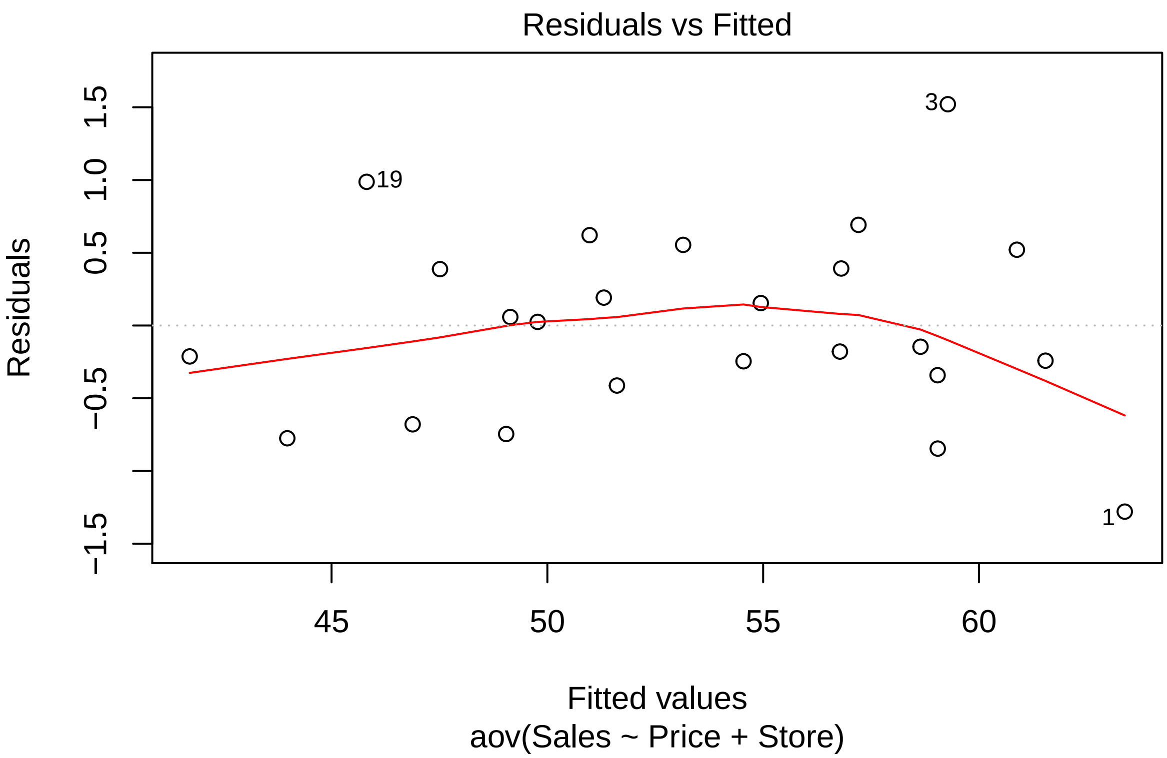
e.
The point estimation of σ^2 is MSE, which is 0.12639.
The point estimation of σ_α^2 is $\frac{MSA-MSA(B)}{nb}$, which is (114.4 - 23.4)/9 = 10.111
The point estimation of $\sigma_{B(A)}^2$ is $\frac{MSA(B)-MSE}{n}$, which is (23.4 - 0.12639)/3 = 7.7579
$\sigma_Y^2 = \sigma_\alpha^2 + \sigma_{B(A)}^2 + \sigma^2$, σ_α^2 is the largest, so σ_α^2 appears to be most important in the total variance.

Q27.6
a.

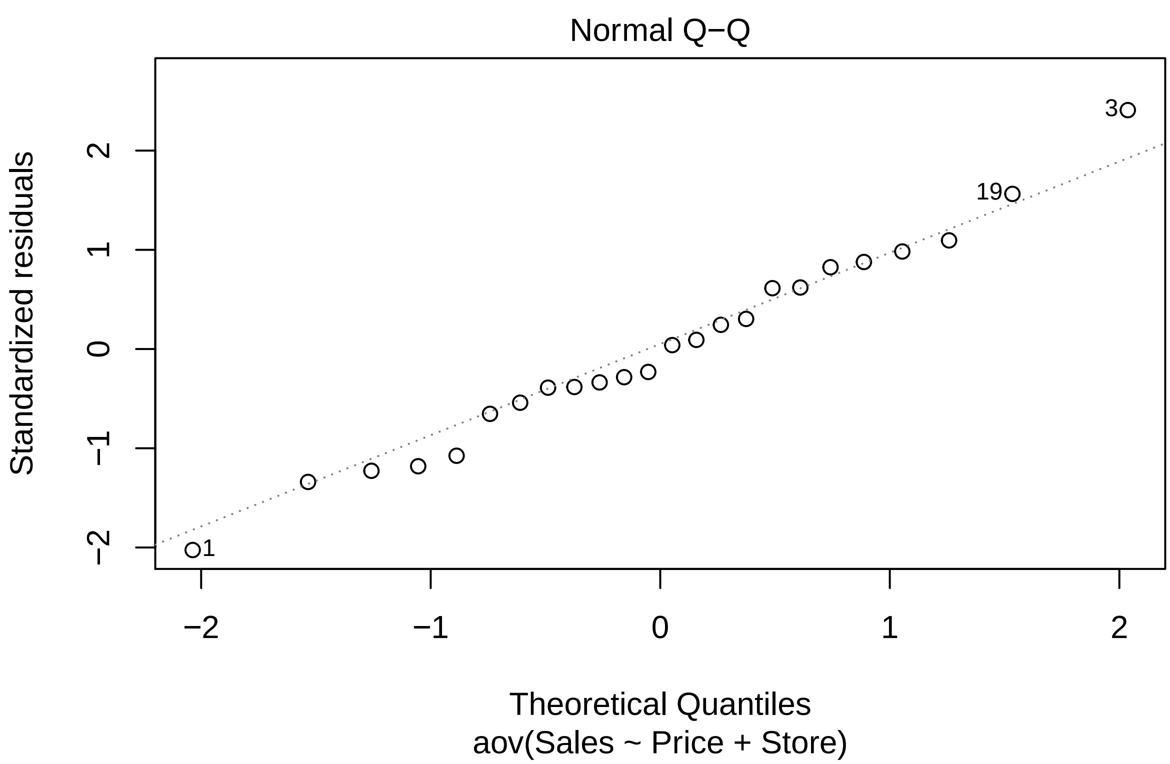
```
data = read.csv("~/academic/Sta207/PR27.6.csv")
data$Store = as.factor(data$Store)
data$Price = as.factor(data$Price)
#aov_out = aov(Sales ~ Price + Error(Store), data = data)
aov_out = aov(Sales ~ Price + Store, data = data)
aov_out$residuals
```

```
##         1         2         3         4         5         6         7
## -1.279167 -0.241667  1.520833 -0.845833  0.691667  0.154167  0.620833
##         8         9        10        11        12        13        14
##  0.058333 -0.679167  0.554167  0.191667 -0.745833  0.520833 -0.341667
##        15        16        17        18        19        20        21
## -0.179167 -0.145833  0.391667 -0.245833  0.987500 -0.775000 -0.212500
##        22        23        24
## -0.412500  0.025000  0.387500
```

```
plot(aov_out, which = 1)
```



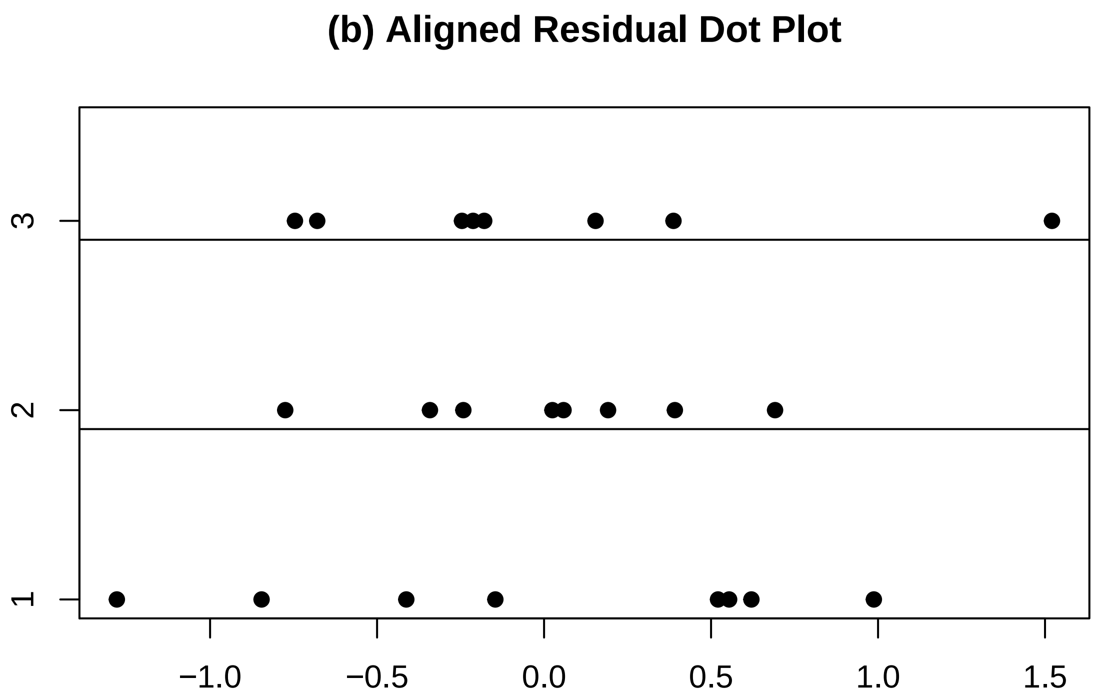

```r
plot(aov_out, which = 2)
```



There is nonlinear pattern in the Residuals vs. Fitted value plot. But is not severe. From the QQ plot, we can find that the residuals are distributed approximately normal. The model seems to be approperiate.

b.

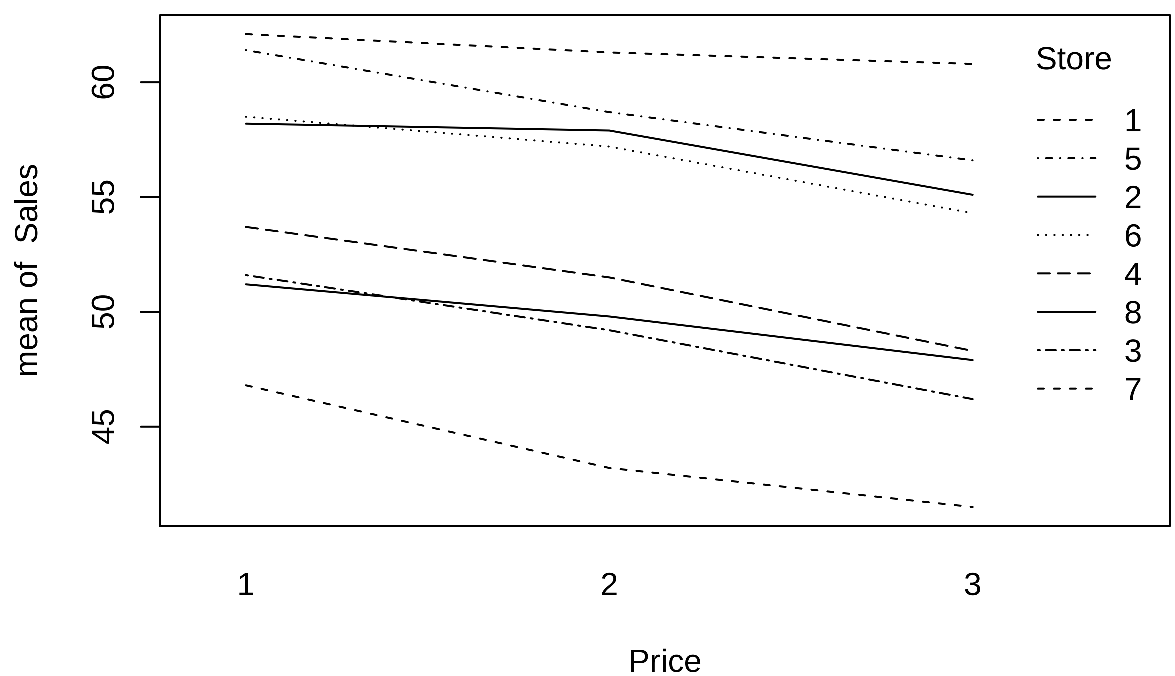
```r
stripchart(split(resid(aov_out), data$Price), method = "stack",  pch = 19)
abline(h = seq(2, 4)-0.1)
title("(b) Aligned Residual Dot Plot")
```



From the aligned residual dot plots by price, we find that the residuals' distribution is similar under different price levels, so the assumption of constancy of the error variance holds.

c.

```r
with(data, interaction.plot(Price, Store, Sales))
```



From the interaction plot, we find that the pattern under different price levels are approximately paralleled to each other, so there seems to be no interaction effect. And the assumption of no interaction appears to be reasonable.

d.

```
#Tukey's test for additivity
alpha = with(data, tapply(Sales,list(Store), mean)) - mean(data$Sales)
beta = with(data, tapply(Sales,list(Price), mean)) - mean(data$Sales)
D = 0
for(i in 1:nrow(data)){
  D = D + data$Sales[i]*alpha[data$Store[i]]*beta[data$Price[i]]
}
D = D/(sum(alpha^2)*sum(beta^2))
SSAB = D^2*sum(alpha^2)*sum(beta^2)
SSAB
```

```
##     1
## 2.941
```

$SSAB^* = 2.940951$
$SSE_{new} = SSE - SSAB^* = 9.6 - 2.940951 = 6.659049$
$H_0 : D = 0 \; vs. \; H_a : D \neq 0$
$F^* = \frac{SSAB^*/1}{SSE_{new}/13} = 5.741415$
If $F^* > F(0.99, 1, 13)$, we reject H_0, otherwise we conclude H_0.
$F^* = 5.741415 < 9.073806 =$ F(0.99, 1, 13), so we conclude H_0, the interaction effect is insignificant.
The P-value is 0.03232378.

Q27.7
a.

```
summary(aov_out)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Price        2     67    33.7    49.4 4.6e-07 ***
## Store        7    745   106.5   155.7 3.5e-12 ***
## Residuals   14     10     0.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b.
$H_0 : \tau_j = 0 \; vs. \; H_a : exist \; \tau_j \neq 0$
$F^* = MSTR/MSE = 33.74/0.68 = 49.35$
If $F^* > F(0.95, 2, 14)$, we reject H_0, otherwise we conclude H_0.
$F^* = 49.35 > 3.738892 = F(0.95, 2, 14)$, so we reject H_0, the mean sales of grapefruits differ for the three price levels.
The P-value is 4.57e-07.

c.
By Tukey procedure: T = $\frac{1}{\sqrt{2}} q(0.95, 3, 14) = 2.61728$
By Bonferroni procedure: B = qt(1 - 0.05/(2*3), 14) = 3.7
By Scheffe procedure: S = $\sqrt{2F(0.95, 2, 14)} = 2.734554$
Obviously, Tukey procedure is the best.
MSE = 0.68

$s(\bar{Y}_{\cdot i} - \bar{Y}_{\cdot j}) = \sqrt{0.68375 * 2/8} = 0.4134459$
for $\hat{L_1} = \bar{Y_{\cdot 1}} - \bar{Y_{\cdot 2}} = 1.8375$, the 95% confidence interval is $1.8375 \pm 2.61728 \times 0.4134459$, which is (0.7553963, 2.919604).
for $\hat{L_2} = \bar{Y_{\cdot 1}} - \bar{Y_{\cdot 3}} = 4.1$, the 95% confidence interval is $4.1 \pm 2.61728 \times 0.4134459$, which is (3.017896, 5.182104).
for $\hat{L_3} = \bar{Y_{\cdot 2}} - \bar{Y_{\cdot 3}} = 2.2625$, the 95% confidence interval is $4.1 \pm 2.61728 \times 0.4134459$, which is (1.180396, 3.344604).

d.

```
nb = 8
r = 3
MSBL = 106.45500
MSBLTR = 0.68375
#According to our book, unbiased estimator
sr2_unbiased = ((nb-1)*MSBL + (nb)*(r-1)*MSBLTR)/(nb*r-1)
E_unbiased = sr2_unbiased/MSBLTR
E_unbiased
```

```
## [1] 48.08
```

So $\hat{E} = 48.08043876$, which is quite large, so the repeated measures desgin is very effective compared to a completely randomized design.