ADVANCED DATA ANALYSIS

HW3

Fan Yang
UNI: fy2232
02/15/2018

Problem 1

(Data in file softdrink.txt) A soft-drink manufacturer uses five agents (1, 2, 3, 4, 5) to handle the premium distributions for its various products. The marketing director desired to study the timeliness with which the premiums are distributed. Twenty transactions for each agent were selected at random and the time lapse (in days) for handling each transaction was determined. Assume the one way anova model is appropriate.

Load data 'softdrink.txt':

In [1]:

```
softdrink <- read.table("sofdrink.txt", header = T)
head(softdrink,5)
softdrink$Agent <- as.factor(softdrink$Agent)
```

Time_lapse	Agent	observation
24	1	1
24	1	2
29	1	3
20	1	4
21	1	5

(a)

(2pt) Obtain the analysis of variance table and test whether or not the mean time lapse differs for the five agents. Use $\alpha = 0.05$.

In [2]:

```
anova.1a <- aov(Time_lapse~Agent,data=softdrink)
summary(anova.1a)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)    
Agent        4   4430  1107.5   147.2 <2e-16 ***
Residuals   95    715     7.5                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the anova table above, we get the p-value is less than 2e-16, under $\alpha = 0.05$, we can reject H_0 and conclude that the mean time lapse differs for the five agents.

(b)

(2pt) Test for all pairs of factor level means whether or not they differ using the Tukey procedure with $\alpha = 0.05$***. Set up groups of factor levels whose means do not differ. Use a paired comparison plot to summarize the results.***

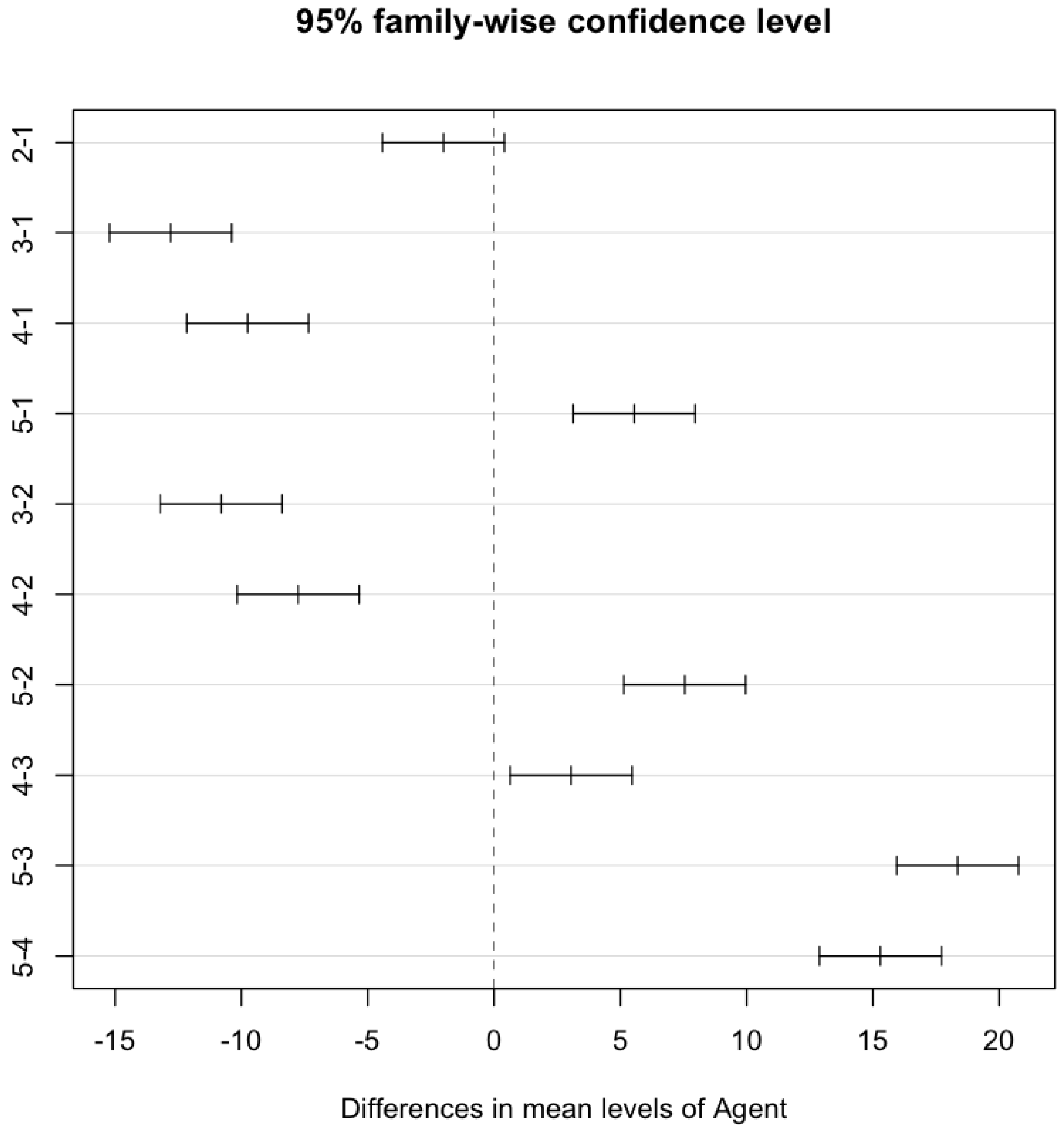
```
In [3]: tk.1b <- TukeyHSD(anova.1a,"Agent")
        tk.1b
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Time_lapse ~ Agent, data = softdrink)

$Agent
      diff         lwr         upr     p adj
2-1  -2.00  -4.4119302   0.4119302 0.1520498
3-1 -12.80 -15.2119302 -10.3880698 0.0000000
4-1  -9.75 -12.1619302  -7.3380698 0.0000000
5-1   5.55   3.1380698   7.9619302 0.0000001
3-2 -10.80 -13.2119302  -8.3880698 0.0000000
4-2  -7.75 -10.1619302  -5.3380698 0.0000000
5-2   7.55   5.1380698   9.9619302 0.0000000
4-3   3.05   0.6380698   5.4619302 0.0059245
5-3  18.35  15.9380698  20.7619302 0.0000000
5-4  15.30  12.8880698  17.7119302 0.0000000
```

```
In [4]: plot(tk.1b)
```



Under the Tukey procedure with $\alpha = 0.05$, there are only two levels have the same mean time lapse, which is $\{1, 2\}$. While the rest levels $\{3\}, \{4\}, \{5\}$ all differ from each other.

(c)

(2pt)The marketing director wishes to compare the mean time lapses for agents 1, 3 and 5. Obtain the pairwise confidence interval for all pairwise comparisons among these three treatment means using the Bonferroni procedure with a 90% family confidence coefficient. Interpret your result.

```
In [5]: ind = (softdrink$Agent != 2)&(softdrink$Agent!=4)
        softdrink.sel = softdrink[ind,]
        pairwise.t.test(softdrink.sel[,1],softdrink.sel[,2],
                        pool.sd=TRUE, p.adjust.method="bonf" )
```

```
        Pairwise comparisons using t tests with pooled SD

data:  softdrink.sel[, 1] and softdrink.sel[, 2]

  1       3
3 < 2e-16 -
5 8.4e-08 < 2e-16

P value adjustment method: bonferroni
```

Under the Bonferroni procedure, we can reject H_0 for every pair comparison, and conclude that every pair of level differs.

```
In [6]: anova.1c <- aov(softdrink.sel$Time_lapse~softdrink.sel$Agent,
                        contrasts=contrasts(softdrink.sel$Agent))
        summary.lm(anova.1c)
```

```
Call:
aov(formula = softdrink.sel$Time_lapse ~ softdrink.sel$Agent,
    contrasts = contrasts(softdrink.sel$Agent))

Residuals:
   Min     1Q Median     3Q    Max
-8.100 -1.550 -0.325  1.988  6.250

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           24.5500     0.6102  40.231  < 2e-16 ***
softdrink.sel$Agent3 -12.8000     0.8630 -14.832  < 2e-16 ***
softdrink.sel$Agent5   5.5500     0.8630   6.431  2.8e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.729 on 57 degrees of freedom
Multiple R-squared:  0.893,     Adjusted R-squared:  0.8892
F-statistic: 237.8 on 2 and 57 DF,  p-value: < 2.2e-16
```

A $100(1-\alpha)\%$ confidence interval for $L = \sum c_i \mu_i$ is

$$\hat{L} \pm t_{n-k}(\alpha/2) SE(\hat{L})$$

where $SE(\hat{L}) = 0.8630$ according to the above table.

In [7]:

```
SE = 0.8630
L31 <- mean(softdrink.sel$Time_lapse[softdrink.sel$Agent==3] -
    softdrink.sel$Time_lapse[softdrink.sel$Agent==1])
L51 <- mean(softdrink.sel$Time_lapse[softdrink.sel$Agent==5] -
    softdrink.sel$Time_lapse[softdrink.sel$Agent==1])
L53 <- mean(softdrink.sel$Time_lapse[softdrink.sel$Agent==5] -
    softdrink.sel$Time_lapse[softdrink.sel$Agent==3])
L31;L51;L53
```

```
-12.8

5.55

18.35
```

Under the Bonferroni procedure with $1 - \alpha = 90\%$ family confidence coefficient, we have every pair confidence coefficient equals to $1 - \frac{0.01}{3}$

In [8]:

```
tn.k <- qt(1-0.01/2/3,60-2)
L31 + tn.k*SE
L31 - tn.k*SE
L51 + tn.k*SE
L51 - tn.k*SE
L53 + tn.k*SE
L53 - tn.k*SE
```

```
-10.1578025096169

-15.4421974903831

8.19219749038313

2.90780250961687

20.9921974903831

15.7078025096169
```

Under 90% family confidence coefficient, each level can be set to $(1 - 0.10/3)$. Compute mean for each levels and get the confidence intervals.

	confidence interval	
3-1 :	[-15.4421,-10.1578]	
5-1 :	[2.90780,8.19220]	
5-3 :	[15.7078,20.9921]	

There is no intervals that intesects with 0, therefore the means differ from each other.

(d)

(2pt) Agents 1 and 2 distribute merchandise only, agents 3 and 4 distribute cash value coupons only and agent 5 distributes both merchandise and coupons. Estimate the contrast

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$

using a 95% confidence interval. Interpret your result.

```
In [9]: anova.1d <- aov(softdrink$Time_lapse~softdrink$Agent,
                        contrasts=contrasts(softdrink$Agent))
        summary.lm(anova.1d)
```

```
Call:
aov(formula = softdrink$Time_lapse ~ softdrink$Agent, contrasts = contr
asts(softdrink$Agent))

Residuals:
   Min     1Q Median     3Q    Max 
-8.100 -1.762 -0.325  1.975  6.450 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       24.5500     0.6133  40.030  < 2e-16 ***
softdrink$Agent2  -2.0000     0.8673  -2.306   0.0233 *  
softdrink$Agent3 -12.8000     0.8673 -14.758  < 2e-16 ***
softdrink$Agent4  -9.7500     0.8673 -11.241  < 2e-16 ***
softdrink$Agent5   5.5500     0.8673   6.399 5.88e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.743 on 95 degrees of freedom
Multiple R-squared:  0.8611,    Adjusted R-squared:  0.8552 
F-statistic: 147.2 on 4 and 95 DF,  p-value: < 2.2e-16
```

A $100(1-\alpha)\%$ confidence interval for $L = \sum c_i \mu_i$ is

$$\hat{L} \pm t_{n-k}(\alpha/2) SE(\hat{L})$$

where $SE(\hat{L}) = 0.8673$ according to the above table.

```
In [10]: SE = 0.8673
         L <- mean((softdrink$Time_lapse[softdrink$Agent==1]+
                   softdrink$Time_lapse[softdrink$Agent==2]) / 2 -
                   (softdrink$Time_lapse[softdrink$Agent==3] +
                   softdrink$Time_lapse[softdrink$Agent==4]) / 2 )
         L
```

10.275

```
In [11]: tn.k <- qt(1-0.05/2,100-4)
         L + tn.k*SE
         L - tn.k*SE
```

```
11.9965768933834

8.55342310661657
```

The contrast

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$

has a 95% confidence interval of [8.5534,11.9966].

(e)

(2pt) Estimate the following comparisons with a 95% confidence interval using the Bonferroni method

$$L_1 = \mu_1 - \mu_2, L_2 = \frac{\mu_1 + \mu_2}{2} - \mu_5, L_3 = \frac{\mu_3 + \mu_4}{2} - \mu_5.$$

A $100(1-\alpha)\%$ confidence interval for $L = \sum c_i \mu_i$ is

$$\hat{L} \pm t_{n-k}(\alpha/2) SE(\hat{L})$$

where $SE(\hat{L}) = 0.8673$ according to the above table.

```
In [12]: SE = 0.8673
         L1 <- mean(softdrink$Time_lapse[softdrink$Agent==1] -
                    softdrink$Time_lapse[softdrink$Agent==2])
         L2 <- mean(softdrink$Time_lapse[softdrink$Agent==1]/2 +
                    softdrink$Time_lapse[softdrink$Agent==2]/2 -
                    softdrink$Time_lapse[softdrink$Agent==5])
         L3 <- mean(softdrink$Time_lapse[softdrink$Agent==3]/2 +
                    softdrink$Time_lapse[softdrink$Agent==4]/2 -
                    softdrink$Time_lapse[softdrink$Agent==5])
         L1;L2;L3
```

```
2

-6.55

-16.825
```

Under the Bonferroni procedure with $1 - \alpha = 95\%$ family confidence coefficient, we have every pair confidence coefficient equals to $1 - \frac{0.05}{3}$.

In [13]:

```
tn.k1 <- qt(1-0.05/2/3,100-2)
tn.k2 <- qt(1-0.05/2/3,100-3)
tn.k3 <- qt(1-0.05/2/3,100-3)
L1 - tn.k1*SE
L1 + tn.k1*SE
L2 - tn.k2*SE
L2 + tn.k2*SE
L3 - tn.k3*SE
L3 + tn.k3*SE
```

```
-0.112541845523311

4.11254184552331

-8.66292171474831

-4.43707828525169

-18.9379217147483

-14.7120782852517
```

Under 90% family confidence coefficient, each level can be set to $(1-0.05/3)$. Compute mean for each levels and get the confidence intervals.

	confidence interval	
L_1 :	[-0.1125, 4.1125]	
L_2 :	[-8.6629, -4.4371]	
L_3:	[-18.9379, -14.7121]	

(f)

(2pt) Of all premium distributions, 25% are handled by agent 1, 20% are handled by agent 2, 20% are handled by agent 3 , 20% are handled by agent 4 and 15% are handled by agent 5. Estimate the overall mean time lapse for premium distributions with a 95% confidence interval.

```
In [14]: anova.1f <- aov(softdrink$Time_lapse~softdrink$Agent,
                         contrasts=contrasts(softdrink$Agent))
         summary.lm(anova.1f)
```

```
Call:
aov(formula = softdrink$Time_lapse ~ softdrink$Agent, contrasts = contr
asts(softdrink$Agent))

Residuals:
   Min     1Q Median     3Q    Max 
-8.100 -1.762 -0.325  1.975  6.450 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       24.5500     0.6133  40.030  < 2e-16 ***
softdrink$Agent2  -2.0000     0.8673  -2.306   0.0233 *  
softdrink$Agent3 -12.8000     0.8673 -14.758  < 2e-16 ***
softdrink$Agent4  -9.7500     0.8673 -11.241  < 2e-16 ***
softdrink$Agent5   5.5500     0.8673   6.399 5.88e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.743 on 95 degrees of freedom
Multiple R-squared:  0.8611,    Adjusted R-squared:  0.8552 
F-statistic: 147.2 on 4 and 95 DF,  p-value: < 2.2e-16
```

A $100(1-\alpha)\%$ confidence interval for $L = \sum c_i \mu_i$ is

$$\hat{L} \pm t_{n-k}(\alpha/2) SE(\hat{L})$$

where $SE(\hat{L}) = 0.8673$ according to the above table.

```
In [15]: SE.1f = 0.8673
         L.1f <- mean(0.25*softdrink$Time_lapse[softdrink$Agent==1] +
                      0.20*softdrink$Time_lapse[softdrink$Agent==2] +
                      0.20*softdrink$Time_lapse[softdrink$Agent==3] +
                      0.20*softdrink$Time_lapse[softdrink$Agent==4] +
                      0.15*softdrink$Time_lapse[softdrink$Agent==5] )
         L.1f
```

20.4725

```
In [16]: tn.k <- qt(1-0.05/2,100-5)
         L.1f - tn.k*SE.1f
         L.1f + tn.k*SE.1f
```

18.7506918046597

22.1943081953403

From the output above, the overall mean time lapse for premium distributions with a 95% confidence interval is [18.7507,22.1943].

Problem 2

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amount of the two active ingredients (factors A and B) in the compound were varied are three levels each. Randomization was used in assigning four volunteers the the nine treatments. The data on hours of relief is in the file (hay.tx).

Load data 'Hayfever.txt':

In [17]:

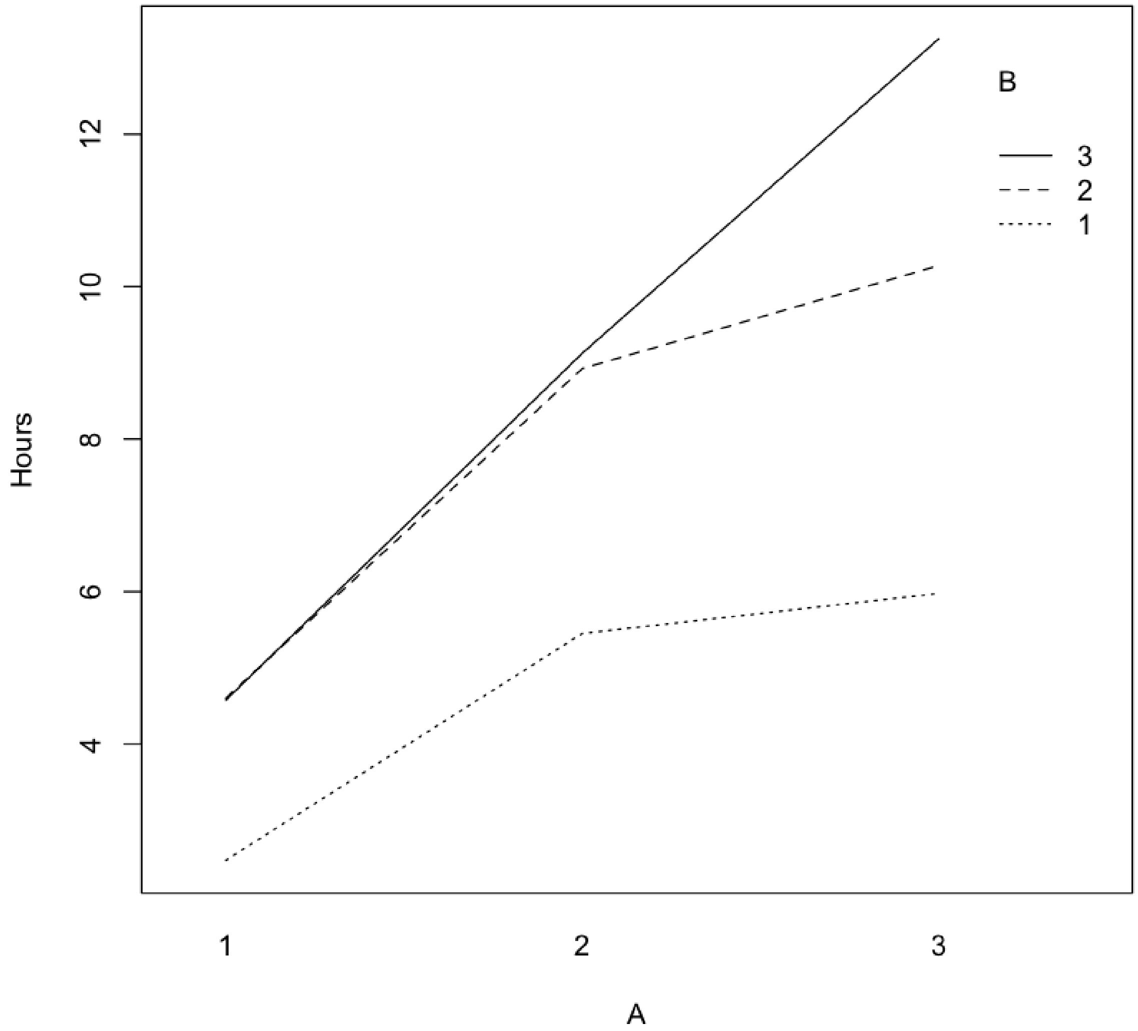
```
Hayfever <- read.table("Hayfever.txt",header = T)
Hayfever$A <- as.factor(Hayfever$A)
Hayfever$B <- as.factor(Hayfever$B)
head(Hayfever,5)
```

Hours	A	B	Observation
2.4	1	1	1
2.7	1	1	2
2.3	1	1	3
2.5	1	1	4
4.6	1	2	1

(a)

(2pt) Construct an interaction plot. Does it suggest that that there is an interaction between A and B? Test whether or not the two factor interact using $\alpha = 0.05$. What is the p-value of this test?

```
In [18]: interaction.plot(Hayfever$A,Hayfever$B,Hayfever$Hours,
                          trace.label="B", xlab="A", ylab = "Hours")
```



In the interaction plot, we can say that there is an interaction between A and B.

```
In [19]: summary(aov(Hours~A*B,data=Hayfever))

                Df Sum Sq Mean Sq F value Pr(>F)
A                2 220.02  110.01  1827.9 <2e-16 ***
B                2 123.66   61.83  1027.3 <2e-16 ***
A:B              4  29.43    7.36   122.2 <2e-16 ***
Residuals       27   1.63    0.06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value of the test for interaction is less than 2e-16. Under $\alpha = 0.05$, We can reject H_0 and conclude that there is interaction.

(b)

(2pt) Fit the model without the interaction and test whether the effects of the two factors are present. Use $\alpha = 0.05$.

```
In [20]: summary(aov(Hours~A+B,data=Hayfever))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
A            2 220.02  110.01  109.83 8.51e-15 ***
B            2 123.66   61.83   61.73 1.55e-11 ***
Residuals   31  31.05    1.00                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From this output and under $\alpha = 0.05$, we can conclude that A level means (p-value=8.51e-15) are different and B level means (p-value=1.55e-11) are also different (There an both A and B effects)

(c)

(2pt) Assume that the model with the interaction is appropriate and construct a 95% confidence interval for μ_{23}.

```
In [21]: lm.2c <- lm(Hours~A*B,data=Hayfever)
         newdata23 <- data.frame(A='2',B='3')
         predict(lm.2c,newdata23,interval="confidence",level=0.95)
```

	fit	lwr	upr
1	9.125	8.873316	9.376684

See from the above results, the 95% confidence interval for μ_{23} is [8.873316, 9.376684].

(d)

(2pt) Assume that the model with the interaction is appropriate, estimate $L = \mu_{12} - \mu_{11}$ with a 95% confidence interval. Interpret your result.

In [22]:
```
anova.2d<-aov(Hours~A*B,contrasts=list(A= contr.sum,B=contr.sum),
              data=Hayfever)
summary.lm(anova.2d)
```

```
Call:
aov(formula = Hours ~ A * B, data = Hayfever, contrasts = list(A = cont
r.sum,
    B = contr.sum))

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4250 -0.1750  0.0125  0.1875  0.3500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.18333    0.04089 175.684  < 2e-16 ***
A1          -3.30000    0.05782 -57.070  < 2e-16 ***
A2           0.65000    0.05782  11.241 1.09e-11 ***
B1          -2.55000    0.05782 -44.099  < 2e-16 ***
B2           0.75000    0.05782  12.970 4.10e-13 ***
A1:B1        1.14167    0.08178  13.961 7.22e-14 ***
A2:B1        0.16667    0.08178   2.038 0.051446 .  
A1:B2       -0.03333    0.08178  -0.408 0.686767    
A2:B2        0.34167    0.08178   4.178 0.000276 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2453 on 27 degrees of freedom
Multiple R-squared:  0.9957,    Adjusted R-squared:  0.9944 
F-statistic: 774.9 on 8 and 27 DF,  p-value: < 2.2e-16
```

A $100(1-\alpha)\%$ confidence interval for $L = \sum c_i \mu_i$ is

$$\hat{L} \pm t_{n-k}(\alpha/2) SE(\hat{L})$$

where $SE(\hat{L}) = 0.08178$ according to the above table.

In [23]:
```
SE.2d = 0.08178
L.2d <- mean(Hayfever$Hours[Hayfever$A==1&Hayfever$B==2] - 
             Hayfever$Hours[Hayfever$A==1&Hayfever$B==1])
```

In [24]:
```
tn.k <- qt(1-0.05/2,36-2)
L.2d + tn.k*SE.2d
L.2d - tn.k*SE.2d
```

2.291196955972

1.958803044028

From the output above, $L = \mu_{12} - \mu_{11}$ has a 95% confidence interval [1.9588,2.2912]. The mean of $L = \mu_{12} - \mu_{11}$ is 2.125.