STA207 homework3

Juanjuan Hu, Zhen Zhang

January 29, 2016

23.13

```
transform_function <- function(x) {
  log10(x + 1)
}
kidney$Days <- transform_function(as.numeric(as.character(kidney$Days)))
kidney <- subset(kidney, Duration != 2 | Weight != 1) # empty y_21 cell
```

(a)

The full model is:

$Y_{ijk} = \mu_{..} + \alpha_1 * X_{ij1} + \beta_1 * X_{ij2} + \beta_2 * X_{ij3} + \epsilon_{ijk}$

The reduced model for testing for factor A main effects is:

$Y_{ijk} = \mu_{..} + \beta_1 * X_{ij2} + \beta_2 * X_{ij3} + \epsilon_{ijk}$

The reduced model for testing for factor B main effects is:

$Y_{ijk} = \mu_{..} + \alpha_1 * X_{ij1} + \epsilon_{ijk}$

Where:

$$X_{ij1} = \begin{cases} 1, & \text{if case from level 1 for factor A} \\ -1, & \text{if case from level 2 for factor A} \end{cases}$$

$$X_{ij2} = \begin{cases} 1, & \text{if case from level 1 for factor B} \\ -1, & \text{if case from level 3 for factor B} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ij3} = \begin{cases} 1, & \text{if case from level 2 for factor B} \\ -1, & \text{if case from level 3 for factor B} \\ 0, & \text{otherwise} \end{cases}$$

(b)

```
kidney$X1 = ifelse(kidney$Duration == 1, 1, -1)
kidney$X2 = ifelse(kidney$Weight == 1,1, ifelse(kidney$Weight == 3, -1, 0))
kidney$X3 = ifelse(kidney$Weight == 2,1, ifelse(kidney$Weight == 3, -1, 0))
fullModel = with(kidney, lm(Days~X1+X2+X3))
fullModel
```

```
##
## Call:
## lm(formula = Days ~ X1 + X2 + X3)
##
## Coefficients:
## (Intercept)           X1           X2           X3
##     0.66939      0.11733     -0.34323      0.02608
```

```
sse.full = anova(fullModel)[4,2]
sse.full
```

```
## [1] 4.489821
```

```
reducedModelTestA = with(kidney, lm(Days~X2+X3))
reducedModelTestA
```

```
##
## Call:
## lm(formula = Days ~ X2 + X3)
##
## Coefficients:
## (Intercept)           X2           X3
##     0.70850     -0.26502     -0.01303
```

```
sse.reduceA = anova(reducedModelTestA)[3,2]
sse.reduceA
```

```
## [1] 5.040447
```

```
reducedModelTestB = with(kidney, lm(Days~X1))
reducedModelTestB
```

```
##
## Call:
## lm(formula = Days ~ X1)
##
## Coefficients:
## (Intercept)           X1
##     0.75520      0.03152
```

```
sse.reduceB = anova(reducedModelTestB)[2,2]
sse.reduceB
```

```
## [1] 7.10425
```

From R output above, we can fit the full model as:

$\hat{Y} = 0.66939 + 0.11733 * X_{ij1} - 0.34323 * X_{ij2} + 0.02608 * X_{ij3}$

And the SSE of full model is 4.4898209.

The reduced model for testing A main effects is:

$\hat{Y} = 0.70850 - 0.26502 * X_{ij2} - 0.01303 * X_{ij3}$

The corresponding SSE is 5.0404474.

The reduced model for testing B main effects is:

$\hat{Y} = 0.75520 + 0.03152 * X_{ij1}$

The corresponding SSE is 7.1042504.

TESTING A MAIN EFFECTS:

$H_0 : \alpha_1 = 0$

$H_1 : \alpha_1 \neq 0$

$F^* = \frac{(5.0404474 - 4.4898209)/1}{(4.4898209)/46} = 0.5506265/0.0976048 = 5.6414$

$F(0.95, 1, 46) = 4.0517$

$p - val = 0.02176$

The decision rule is: if F^* is greater thatn 4.0517, then reject H_0, otherwise, accept H_1. Here, $5.6414 \geq 4.0517$, so we reject H_0, concluding that factor A main effects are present. The p-value is 0.02176, which is less than 0.05, leading to the same conclusion.

TESTING B MAIN EFFECTS:

$H_0 : \beta_1 = \beta_2 = 0$

$H_1 : notall \beta equalto 0$

$F^* = \frac{(7.1042504 - 4.4898209)/2}{(4.4898209)/46} = 1.307215/0.0976048 = 13.39294$

$F(0.95, 2, 46) = 3.1996$

$p - val = 2.608231e - 05$

The decision rule is: if F^* is greater thatn 3.1996, then reject H_0, otherwise, accept H_1. Here, $13.39294 \geq 4.0517$, so we reject H_0, concluding that factor B main effects are present. The p-value is almost zero, which is less than 0.05, leading to the same conclusion.

23.19

(a)

The ANOVA model is:

$Y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}$, Where $i = 1, 2, ..., 5$, $j = 1, 2, 3$

The corresponding regression model is:

$Y_{ij} = \mu_{..} + \rho_1 * X_{ij1} + \rho_2 * X_{ij2} + \rho_3 * X_{ij3} + \rho_4 * X_{ij4} + \tau_1 * X_{ij5} + \tau_2 * X_{ij6} + \epsilon_{ij}$

Where:

$$X_{ij1} = \begin{cases} 1, & \text{if case from block 1} \\ -1, & \text{if case from block 5} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ij2} = \begin{cases} 1, & \text{if case from block 2} \\ -1, & \text{if case from block 5} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ij3} = \begin{cases} 1, & \text{if case from block 3} \\ -1, & \text{if case from block 5} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ij4} = \begin{cases} 1, & \text{if case from block 4} \\ -1, & \text{if case from block 5} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ij5} = \begin{cases} 1, & \text{if case from treatment 1} \\ -1, & \text{if case from treatment 3} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ij6} = \begin{cases} 1, & \text{if case from treatment 2} \\ -1, & \text{if case from treatment 3} \\ 0, & \text{otherwise} \end{cases}$$

(b)

The reduced model for testing for differences in the mean reductions in lipid level for treatment is:

$Y_{ij} = \mu_{..} + \rho_1 * X_{ij1} + \rho_2 * X_{ij2} + \rho_3 * X_{ij3} + \rho_4 * X_{ij4} + \epsilon_{ij}$

(c)

```
Yij = c(0.73, 0.67, 0.15, 0.86, 0.75, 0.21, 0.94, 0.81, 0.26, 1.4, 1.32, 0.75, 1.62, 1.41, 0.78)
obs = data.frame(matrix(Yij, 5,3,2))
rownames(obs) = c("block1", "block2", "block3", "block4", "block5")
names(obs) = c("treatment1", "treatment2", "treatment3")
obs[1,3] = NA
obs[5,1] = NA
obs
```

```
##        treatment1 treatment2 treatment3
## block1       0.73       0.67         NA
## block2       0.86       0.75       0.21
## block3       0.94       0.81       0.26
## block4       1.40       1.32       0.75
## block5         NA       1.41       0.78
```

```
Y = c(0.73, 0.67, 0.86, 0.75, 0.21, 0.94, 0.81, 0.26, 1.4, 1.32, 0.75, 1.41, 0.78)
X1 = c(1,1,0,0,0,0,0,0,0,0,0,-1,-1)
X2 = c(0,0,1,1,1,0,0,0,0,0,0,-1,-1)
X3 = c(0,0,0,0,0,1,1,1,0,0,0,-1,-1)
X4 = c(0,0,0,0,0,0,0,0,1,1,1,-1,-1)
X5 = c(1,0,1,0,-1,1,0,-1,1,0,-1,0,-1)
X6 = c(0,1,0,1,-1,0,1,-1,0,1,-1,1,-1)
df = cbind(Y, X1, X2, X3, X4, X5,X6)
df = data.frame(df)
```

```
fullModel2 = with(df, lm(Y~X1+X2+X3+X4+X5+X6))
fullModel2
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##      0.8294      -0.3361      -0.2227      -0.1594       0.3273
##          X5           X6
##      0.2508       0.1626
```

```
sse.full2 = anova(fullModel2)[7,2]
sse.full2
```

```
## [1] 0.00350582
```

```
reducedModel2 = with(df, lm(Y~X1+X2+X3+X4))
reducedModel2
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##      0.8457      -0.1457      -0.2390      -0.1757       0.3110
```

```
sse.reduce2 = anova(reducedModel2)[5,2]
sse.reduce2
```

```
## [1] 0.9541833
```

From R output above, we can fit the full model as:

$\hat{Y} = 0.8294 - 0.3361 * X_1 - 0.2227 * X_2 - 0.1594 * X_3 + 0.3273 * X_4 + 0.2508 * X_5 + 0.1626 * X_6$

And the SSE of full model is 0.0035058.

The reduced model for testing differences in the treatment is:

$\hat{Y} = 0.8457 - 0.1457 * X_1 - 0.2390 * X_2 - 0.1757 * X_3 + 0.311 * X_4$

The corresponding SSE is 0.9541833.

TESTING TREATMENT EFFECTS:

$H_0 : \tau_1 = \tau_2 = 0$

$H_1 : notallequaltozero$

$F^* = \frac{(0.9541833-0.0035058)/2}{(0.0035058)/6} = 0.4753/0.00058 = 819.48$

$F(0.95, 2, 6) = 5.1433$

The decision rule is: if F^* is greater thatn 5.1433, then reject H_0, otherwise, accept H_1. Here, $819.48 \geq 4.0517$, so we reject H_0, concluding that the mean reductions in lipid level differ for the three diets. The result is the same as obtained in Problem 23.17d.

(d)

```r
vcov(fullModel2)
```

```
##               (Intercept)            X1            X2            X3
## (Intercept)  4.760990e-05  1.298452e-05 -8.656346e-06 -8.656346e-06
## X1           1.298452e-05  2.448509e-04 -5.193808e-05 -5.193808e-05
## X2          -8.656346e-06 -5.193808e-05  1.644706e-04 -3.029721e-05
## X3          -8.656346e-06 -5.193808e-05 -3.029721e-05  1.644706e-04
## X4          -8.656346e-06 -5.193808e-05 -3.029721e-05 -3.029721e-05
## X5           4.328173e-06 -3.524369e-05 -4.328173e-06 -4.328173e-06
## X6          -8.656346e-06 -1.298452e-05  8.656346e-06  8.656346e-06
##                        X4            X5            X6
## (Intercept) -8.656346e-06  4.328173e-06 -8.656346e-06
## X1          -5.193808e-05 -3.524369e-05 -1.298452e-05
## X2          -3.029721e-05 -4.328173e-06  8.656346e-06
## X3          -3.029721e-05 -4.328173e-06  8.656346e-06
## X4           1.644706e-04 -4.328173e-06  8.656346e-06
## X5          -4.328173e-06  1.051128e-04 -4.328173e-05
## X6           8.656346e-06 -4.328173e-05  8.656346e-05
```

Construct: $L = \tau_1 - \tau_3 = 2 * \tau_1 + \tau_2$

$\hat{L} = 2 * \hat{\tau}_1 + \hat{\tau}_2 = 2 * 0.2508 + 0.1626 = 0.6642$

According to the covariance matrix of model coefficients, $s^2\{\hat{\tau}_1\} = 1.051128e - 04$, $s^2\{\hat{\tau}_2\} = 8.656346e - 05$, $s\{\hat{\tau}_1, \hat{\tau}_2\} = -4.328173e - 05$. Therefore:

$s\{\hat{L}\} = sqrt(4 * 1.051128e - 04 + 8.656346e - 05 + 4 * (-4.328173e - 05)) = 0.0182726$

$t(0.99, 6) = 3.142668$

$\hat{L} + t(0.99, 6) * s\{\hat{L}\} = 0.6642 + 3.142668 * 0.0182726 = 0.7216247$ $\hat{L} - t(0.99, 6) * s\{\hat{L}\} = 0.6642 - 3.142668 * 0.0182726 = 0.6067753$

Therefore, the 98% confidence interval for diffrence in diet1 and diet3 is [0.6068, 0.7216]. We can find that the CI does not include zero, indicating that mean reduction in lipid for diet 1 is significantly larger than the reduction for diet3.

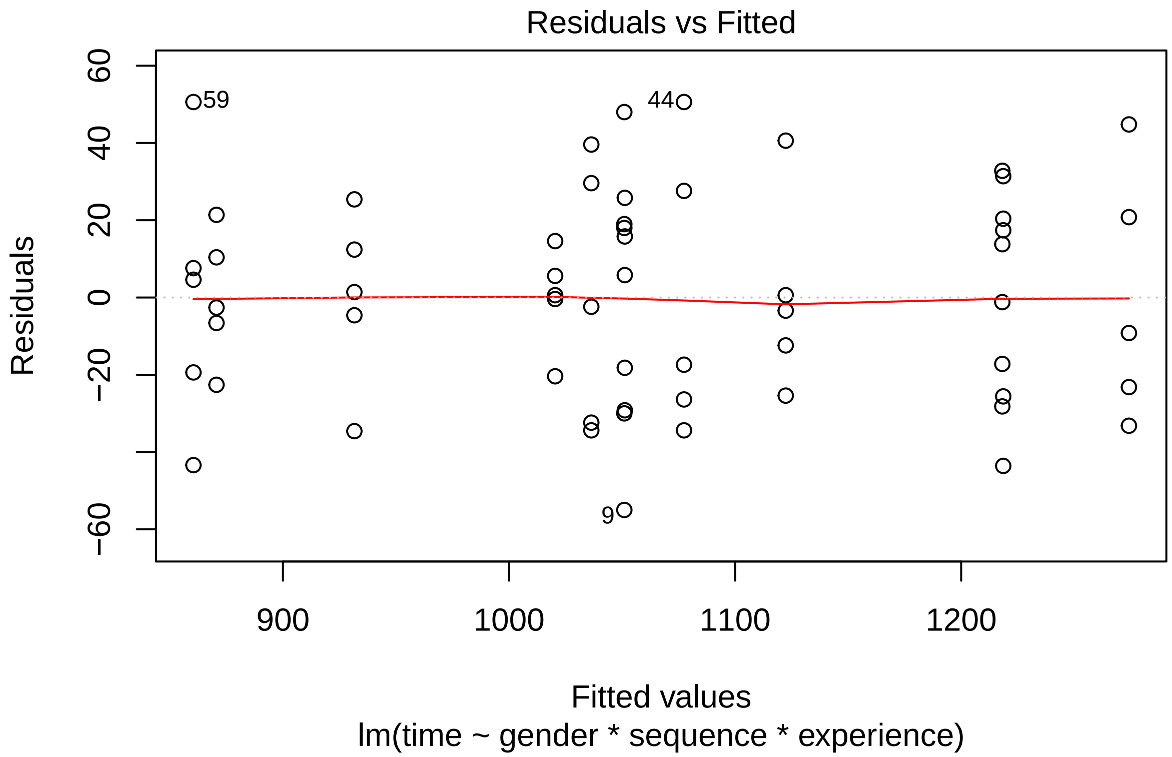
24.12

(d)

```r
assembly <- read.table("CH24PR12.txt")
colnames(assembly) <- c("time", "gender", "sequence", "experience", "replication")
assembly <- assembly[,-5]
assembly$gender <- as.factor(assembly$gender)
assembly$sequence <- as.factor(assembly$sequence)
assembly$experience <- as.factor(assembly$experience)
aov_out <- lm(time ~ gender*sequence*experience, data = assembly)
residuals(aov_out)
```

```
##      1      2      3      4      5      6      7      8      9     10     11     12 
##   31.4  -43.6   17.4   20.4  -25.6  -30.0   48.0   18.0  -55.0   19.0   44.8  -23.2 
##     13     14     15     16     17     18     19     20     21     22     23     24 
##  -33.2   20.8   -9.2   -3.4  -12.4    0.6  -25.4   40.6   -1.2  -28.2  -17.2   13.8 
##     25     26     27     28     29     30     31     32     33     34     35     36 
##   32.8  -18.2   15.8    5.8   25.8  -29.2   29.6   39.6  -32.4  -34.4   -2.4   -6.6 
##     37     38     39     40     41     42     43     44     45     46     47     48 
##  -22.6   10.4   21.4   -2.6   27.6  -34.4  -26.4   50.6  -17.4   -4.6   12.4   25.4 
##     49     50     51     52     53     54     55     56     57     58     59     60 
##  -34.6    1.4    0.6   -0.4   14.6  -20.4    5.6  -19.4    4.6  -43.4   50.6    7.6 
```

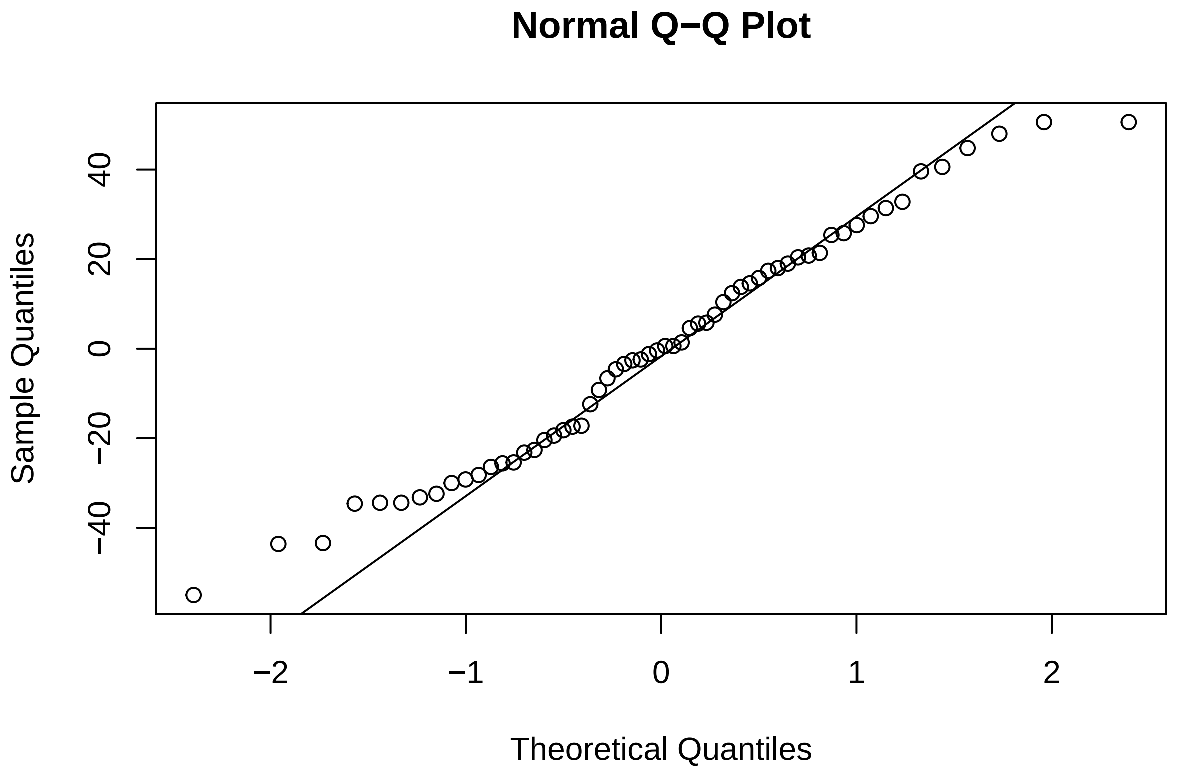
```
plot(aov_out, which = 1)
```



From the plot, we can conclude that this anova model is appropriate.

(e)

```
qqnorm_assembly <- qqnorm(aov_out$residuals)
qqline(aov_out$residuals)
```



```
cor(qqnorm_assembly$x, qqnorm_assembly$y)
```

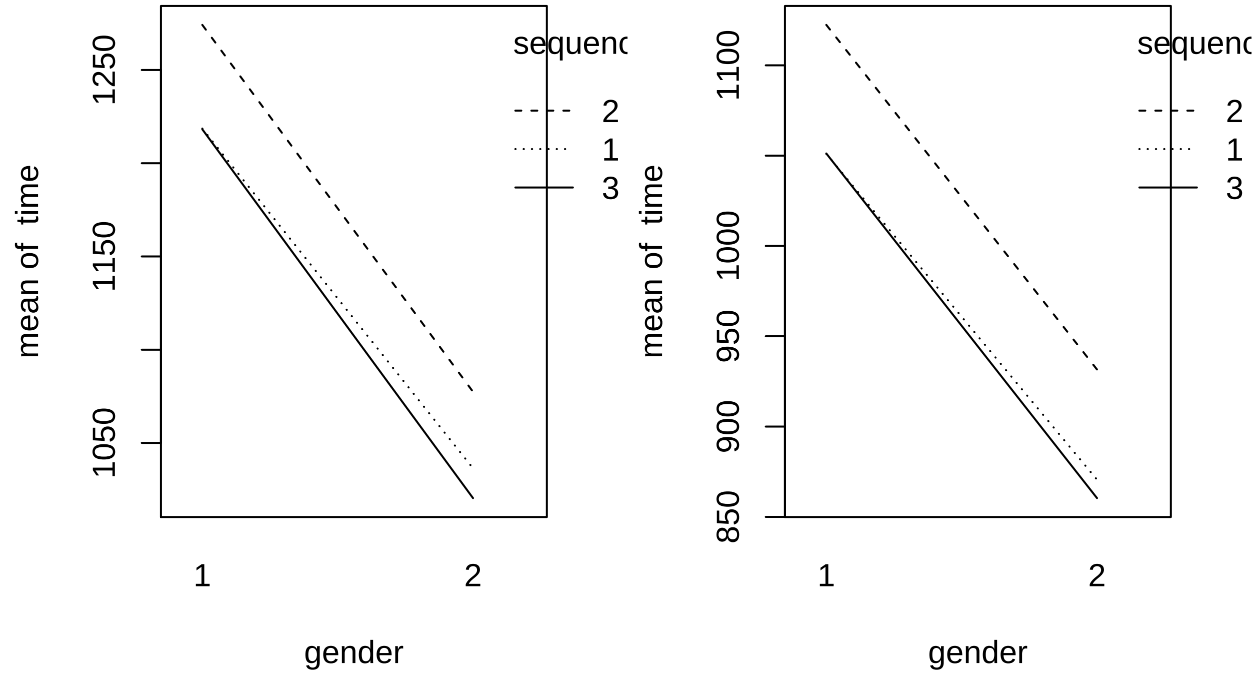
```
## [1] 0.9906479
```

The correlation indicates the normality assumption is reasonable here.

24.13

(a)

```
assembly_experience_split <- split(assembly[, -4], assembly$experience)
par(mfrow = c(1, 2))
with(assembly_experience_split[[1]], interaction.plot(gender, sequence, time))
with(assembly_experience_split[[2]], interaction.plot(gender, sequence, time))
```



Almost no interaction is seen in this plot for the three factors.

(b)

```
(anova_out <- anova(aov_out))
```

```
## Analysis of Variance Table
##
## Response: time
##                            Df Sum Sq Mean Sq  F value    Pr(>F)
## gender                      1 540361  540361 629.7603 < 2.2e-16 ***
## sequence                    2  49320   24660  28.7396  6.22e-09 ***
## experience                  1 382402  382402 445.6679 < 2.2e-16 ***
## gender:sequence             2    543     271   0.3161    0.7305
## gender:experience           1     91      91   0.1064    0.7457
## sequence:experience         2    911     456   0.5310    0.5914
## gender:sequence:experience  2     19      10   0.0111    0.9890
## Residuals                  48  41186     858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c)

Null hypothesis: the coefficient of three-factor interactions == 0

Alternative hypothesis: the coefficient of three-factor interactions != 0

Test statistics: $F^* = 10/858 = 0.01165501$

Decision rule: $F(0.95; 2, 48) =$ `qf(0.95, 2, 48)` $= 3.190727$, then reject H_0 is $F^* > F(0.95; 2, 48)$

Conclusion: since $F^* < F(0.95; 2, 48)$, then we cannot reject H_0.

p-value: 0.9890

(d)

AB-interaction:

Null hypothesis: the coefficient of AB-interactions == 0

Alternative hypothesis: the coefficient of AB-interactions != 0

Test statistics: $F^* = 271/858 = 0.3158508$

Decision rule: $F(0.95; 2, 48) =$ `qf(0.95, 2, 48)` $= 3.190727$, then reject H_0 is $F^* > F(0.95; 2, 48)$

Conclusion: since $F^* < F(0.95; 2, 48)$, then we cannot reject H_0.

p-value: 0.7305

AC-interaction:

Null hypothesis: the coefficient of AC-interactions == 0

Alternative hypothesis: the coefficient of AC-interactions != 0

Test statistics: $F^* = 91/858 = 0.1060606$

Decision rule: $F(0.95; 1, 48) =$ `qf(0.95, 1, 48)` $= 4.042652$, then reject H_0 is $F^* > F(0.95; 1, 48)$

Conclusion: since $F^* < F(0.95; 1, 48)$, then we cannot reject H_0.

p-value: 0.7457

BC-interaction:

Null hypothesis: the coefficient of BC-interactions == 0

Alternative hypothesis: the coefficient of BC-interactions != 0

Test statistics: $F^* = 456/858 = 0.5314685$

Decision rule: $F(0.95; 2, 48) =$ `qf(0.95, 2, 48)` $= 3.190727$, then reject H_0 is $F^* > F(0.95; 2, 48)$

Conclusion: since $F^* < F(0.95; 2, 48)$, then we cannot reject H_0.

p-value: 0.5914

(e)

A

Null hypothesis: the coefficient of A == 0

Alternative hypothesis: the coefficient of A != 0

Test statistics: $F^* = 540361/858 = 629.7914$

Decision rule: $F(0.95; 2, 48) =$ `qf(0.95, 1, 48)` $= 4.042652$, then reject H_0 is $F^* > F(0.95; 1, 48)$

Conclusion: since $F^* > F(0.95; 1, 48)$, then we reject H_0.

p-value: 0.0000

B

Null hypothesis: the coefficient of B == 0

Alternative hypothesis: the coefficient of B != 0

Test statistics: $F^* = 24660/858 = 28.74126$

Decision rule: $F(0.95; 2, 48) =$ `qf(0.95, 2, 48)` $= 3.190727$, then reject H_0 is $F^* > F(0.95; 2, 48)$

Conclusion: since $F^* > F(0.95; 2, 48)$, then we reject H_0.

p-value: 0.0000

C

Null hypothesis: the coefficient of C == 0

Alternative hypothesis: the coefficient of C != 0

Test statistics: $F^* = 382402/858 = 445.69$

Decision rule: $F(0.95; 1, 48) =$ `qf(0.95, 1, 48)` $= 4.042652$, then reject H_0 is $F^* > F(0.95; 1, 48)$

Conclusion: since $F^* > F(0.95; 1, 48)$, then we reject H_0.

p-value: 0.0000

(f)

Only the single effect of A, B and C are significant.

The Kimball Inequality shows that the upper bound for the family level of significance for the set of tests is `1-0.95^7` = 0.3016627.

(g)

Yes, because the upper bound for the family level of significance for the set of tests (0.3016627) is still small than the P-values of AB = 0.7305. We cannot reject H_0 of the coefficient of AB-interaction is 0.

24.14

(a)

```r
assembly_mean <- with(assembly, tapply(time, list(gender = gender, sequence = sequence, experience = exp
(d1 <- apply(assembly_mean, 1, mean)[1] - apply(assembly_mean, 1, mean)[2])
```

```
##     1
## 189.8
```

```r
(d2 <- apply(assembly_mean, 2, mean)[1] - apply(assembly_mean, 2, mean)[2])
```

```
##      1
## -57.25
```

```r
(d3 <- apply(assembly_mean, 2, mean)[1] - apply(assembly_mean, 2, mean)[3])
```

```
##   1
## 6.6
```

```r
(d4 <- apply(assembly_mean, 2, mean)[2] - apply(assembly_mean, 2, mean)[3])
```

```
##     2
## 63.85
```

```r
(d5 <- apply(assembly_mean, 3, mean)[1] - apply(assembly_mean, 3, mean)[2])
```

```
##        1
## 159.6667
```

```r
(mse <- anova_out$`Mean Sq`[8])
```

```
## [1] 858.0417
```

```r
(sd1 <- sqrt(2* mse / (5*2*3)))
```

```
## [1] 7.563252
```

```r
(sd2 <- sqrt(2* mse / (5*2*2)))
```

```
## [1] 9.263054
```

```r
(sd3 <- sd2)
```

```
## [1] 9.263054
```

```r
(sd4 <- sd2)
```

```
## [1] 9.263054
```

```r
(sd5 <- sqrt(2* mse / (5*2*3)))
```

```
## [1] 7.563252
```

Bonferroni: $B = t(1 - \frac{0.05}{5}, 48) = t(0.99, 48) = 2.306 = 2.406581$

So the intervals:

D_1: $189.8 \pm 2.406581 * 7.563252 =$ (171.5984, 208.0016)

D_2: $-57.25 \pm 2.406581 * 9.263054 =$ (-79.54229, -34.95771)

D_3: $6.6 \pm 2.406581 * 9.263054 =$ (-15.69229, 28.89229)

D_4: $63.85 \pm 2.406581 * 9.263054 =$ (41.55771, 86.14229)

D_5: $159.6667 \pm 2.406581 * 7.563252 =$ (141.4651, 177.8683)

(b)

```
(mu231 <- assembly_mean[2,3,1])
```

```
## [1] 1020.4
```

```
(sd231 <- (mse/5)^0.5)
```

```
## [1] 13.09994
```

```
qt(0.975, 48)
```

```
## [1] 2.010635
```

So the interval is $1020.4 \pm 2.011 * 13.09994 = (994.056, 1046.744)$