HW3_Solution

Version 2, by Ethan Rogers

11/15/2023

Problem #1

You plan to analyze data from an experiment with a 2-factor design, where the first factor has a = 3 levels, the second factor has b = 3 levels, and there are 2 replicate observations for each combination of factor levels.

a)

Write out the cell means model, and state the assumptions

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$

b)

Provide the associated ANOVA table with columns Source, Some of Squares, and Degrees of Freedom, with the name or equation in each cell. State the null hypothesis that can be tested with this table

Table 1: ANOVA table

Source	SS	DF
Model	$\text{SSR} = n\sum\sum(\bar{Y}_{ij.} - \bar{Y}_{...})^2$	ab-1
Error	$\text{SSE} = \sum\sum\sum(Y_{ijk} - \bar{Y}_{ij.})^2$	ab(n-1)
Total	$\text{SSTO} = \sum\sum\sum(Y_{ijk} - \bar{Y}_{...})^2$	abn-1

H_0: There is no difference between cell means.

H_a: At least one cell mean is significantly different.

c)

Re-state the cell means model as a linear regression approach, and spell out the vector Y, the design matrix X, the vector of parameters μ, and vector of errors ϵ.

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \\ Y_{311} \\ Y_{312} \\ Y_{321} \\ Y_{322} \\ Y_{331} \\ Y_{332} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \mu 11 \\ \mu 12 \\ \mu 13 \\ \mu 21 \\ \mu 22 \\ \mu 23 \\ \mu 31 \\ \mu 32 \\ \mu 33 \end{bmatrix} + \begin{bmatrix} \xi_{111} \\ \xi_{112} \\ \xi_{121} \\ \xi_{122} \\ \xi_{131} \\ \xi_{132} \\ \xi_{211} \\ \xi_{212} \\ \xi_{221} \\ \xi_{222} \\ \xi_{231} \\ \xi_{232} \\ \xi_{311} \\ \xi_{312} \\ \xi_{321} \\ \xi_{322} \\ \xi_{331} \\ \xi_{332} \end{bmatrix}$$

d)

Write the vector of coefficients C associated with the parameters of the linear regression for the contrast $L = \mu_{12} - \mu_{13}$.

$$L = \mu_{12} - \mu_{13}$$
$$= 0 \cdot \mu_{11} + 1 \cdot \mu_{12} + (-1) \cdot \mu_{13} + 0 \cdot \mu_{21} + 0 \cdot \mu_{21} + 0 \cdot \mu_{21} + 0 \cdot \mu_{31} + 0 \cdot \mu_{32} + 0 \cdot \mu_{33}$$

$$C = (0, 1, -1, 0, 0, 0, 0, 0, 0)$$

e)

Re-write this model as a two-way factor effects model with zero sum constraints, and state the assumptions. Be sure to specify the distributional assumptions and the constraints.

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$

where

$$\sum \alpha_i = 0, \sum \beta_j = 0, \sum_i (\alpha\beta)_{ij} = 0, \sum_j (\alpha\beta)_{ij} = 0,$$

Consequence of the constraints:

$$\alpha_3 = -\alpha_1 - \alpha_2$$

$$\beta_3 = -\beta_1 - \beta_2$$

$$(\alpha\beta)_{3j} = -(\alpha\beta)_{1j} - (\alpha\beta)_{2j} \text{ for all } j$$

$$(\alpha\beta)_{i3} = -(\alpha\beta)_{i1} - (\alpha\beta)_{i2} \text{ for all } i$$

Table 2: ANOVA table

Source	SS	DF
Factor A	$\text{SSA} = nb\sum(\bar{Y}_{i..} - \bar{Y}_{...})^2$	a-1
Factor B	$\text{SSB} = na\sum(\bar{Y}_{.j.} - \bar{Y}_{...})^2$	b-1
AB interactions	$\text{SSAB} = n\sum\sum(\bar{Y}_{ij.} - \bar{Y}_{j..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$	(a-1)(b-1)
Error	$\text{SSE} = \sum\sum\sum(Y_{ijk} - \bar{Y}_{ij.})^2$	ab(n-1)
Total	$\text{SSTO} = \sum\sum\sum(Y_{ijk} - \bar{Y}_{...})^2$	nab-1

f)

Provide the associated ANOVA table with columns Source, Some of Squares, and Degrees of Freedom, with the name or equation in each cell. State the null hypotheses that can be tested with this table.

H_{0a}: Factor A does not have a significant effect on the response. H_{0b}: Factor B does not have a significant effect on the response. H_{0ab}: The interaction of Factor A and Factor B does not have a significant effect on the response.

g)

Re-state the factor effects model as a linear regression approach, and spell out the vector Y, the design matrix X, the vector of parameters μ, and vector of errors ϵ.

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \\ Y_{311} \\ Y_{312} \\ Y_{321} \\ Y_{322} \\ Y_{331} \\ Y_{332} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \end{bmatrix} + \begin{bmatrix} \xi_{111} \\ \xi_{112} \\ \xi_{121} \\ \xi_{122} \\ \xi_{131} \\ \xi_{132} \\ \xi_{211} \\ \xi_{212} \\ \xi_{221} \\ \xi_{222} \\ \xi_{231} \\ \xi_{232} \\ \xi_{311} \\ \xi_{312} \\ \xi_{321} \\ \xi_{322} \\ \xi_{331} \\ \xi_{332} \end{bmatrix}$$

h)

Write the vector of coefficients C *associated with the parameters of the linear regression for the contrast* $L = \mu_{12} - \mu_{13}$.

$$\begin{aligned} L = \mu_{12} - \mu_{13} &= [\mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}] - [\mu + \alpha_1 - \beta_1 - \beta_2 - (\alpha\beta)_{11} - (\alpha\beta)_{12}] \\ &= \beta_1 + 2 \cdot \beta_2 + (\alpha\beta)_{11} + 2 \cdot (\alpha\beta)_{12} \end{aligned}$$

$$C = (0, 0, 0, 1, 2, 1, 2, 0, 0)$$

i)

Re-write this model as a two-way factor effects model with reference (one-hot) constraints, and state the assumptions. Be sure to specify the distributional assumptions and the constraints.

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$
$$\text{where } \alpha_1 = \beta_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$$

j)

Re-state the factor effects model as a linear regression approach, and spell out the vector Y, the design matrix X, the vector of parameters μ, and vector of errors ϵ.

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \\ Y_{311} \\ Y_{312} \\ Y_{321} \\ Y_{322} \\ Y_{331} \\ Y_{332} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \beta_2 \\ \beta_3 \\ (\alpha\beta)_{22} \\ (\alpha\beta)_{23} \\ (\alpha\beta)_{32} \\ (\alpha\beta)_{33} \end{bmatrix} + \begin{bmatrix} \xi_{111} \\ \xi_{112} \\ \xi_{121} \\ \xi_{122} \\ \xi_{131} \\ \xi_{132} \\ \xi_{211} \\ \xi_{212} \\ \xi_{221} \\ \xi_{222} \\ \xi_{231} \\ \xi_{232} \\ \xi_{311} \\ \xi_{312} \\ \xi_{321} \\ \xi_{322} \\ \xi_{331} \\ \xi_{332} \end{bmatrix}$$

k)

Write the vector of coefficients C *associated with the parameters of the linear regression for the contrast* $L = \mu_{12} - \mu_{13}$.

$$\begin{aligned} L = \mu_{12} - \mu_{13} &= [\mu + \beta_2] - [\mu + \beta_3] \\ &= \beta_2 - \beta_3 \end{aligned}$$

$$C = (0, 0, 0, 1, -1, 0, 0, 0, 0)$$

Problem #2

In this question, we will implement the linear regression based estimation of parameters in a 2-factor experimental design above, with a reference constraint (μ_{11} as the reference). The implementations below must be done from scratch (i.e., you cannot use lm or other libraries for linear regression). The implementation does not need to be general (i.e., it's enough to make it work on for this homework). Although linear regression libraries like lm are not allowed in your implementation, you are encouraged to use them to check your work. We recommend you use the dataset from the question 3 to test your implementation.

a)

Implement a linear regression-based estimation of parameters of this model, with a reference constraint (μ_{11} as the reference).

```
data1914<-fread("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%
data1914 <- data1914[order(data1914$V3),]
data1914$V2 <- as.factor(data1914$V2)
data1914$V3 <- as.factor(data1914$V3)

ANOVA33 <- function(data,V2 = V2,V3 = V3){
  data <- data[order(data$V3),]
  data$V2 <- as.factor(data$V2)
  data$V3 <- as.factor(data$V3)
  #get a matrix for every possible rows,
  #you can just input the whole matrix for simplicity
  vector.matrix <- t(matrix(c(1,0,0,0,0,0,0,0,0,
```

```
                          1,1,0,0,0,0,0,0,0,
                          1,0,1,0,0,0,0,0,0,
                          1,0,0,1,0,0,0,0,0,
                          1,1,0,1,0,1,0,0,0,
                          1,0,1,1,0,0,1,0,0,
                          1,0,0,0,1,0,0,0,0,
                          1,1,0,0,1,0,0,1,0,
                          1,0,1,0,1,0,0,0,1),nrow = 9))
  #the index(i,j) for each cell, purpose is identify cells from data
  cell.index <- matrix(c(1,1,1,2,2,2,3,3,3,1,2,3,1,2,3,1,2,3),ncol =2)
  #levels of value for V2 and V3
  level.v2 <- unique(data$V2)
  level.v3 <- unique(data$V3)
  design.matrix <- NULL
  mu.cell <- NULL
  for(i in 1:9){
    #sub set of data for each cell
    sub.data <- data%>%filter(V2 == cell.index[i,2],V3 == cell.index[i,1])
    #mu for each cell
    mu.cell[i] <- mean(sub.data$V1)
    #design matrix for each cell(for lines )
    temp.matrix <- t(matrix(rep(vector.matrix[i,],nrow(sub.data)),ncol = 4))
    design.matrix <- rbind(design.matrix,temp.matrix)
  }
  y <- as.matrix(data$V1)
  beta <- solve(t(design.matrix)%*%design.matrix)%*%t(design.matrix)%*%y
  e <- y-design.matrix%*%beta
  SSE <- sum(e^2)
  SST <- sum((y-mean(y))^2)
  data <- data%>%group_by(V2)%>%summarise(meanA=mean(V1))%>%left_join(data, by="V2")
  data <- data%>%group_by(V3)%>%summarise(meanB=mean(V1))%>%left_join(data, by="V3")
  data <- data%>%mutate(meanAll=mean(V1))
  SSA <- sum(data%>%mutate("SSA"=(meanAll-meanA)^2)%>%dplyr::select(SSA))
  SSB <- sum(data%>%mutate("SSB"=(meanAll-meanB)^2)%>%dplyr::select(SSB))
  SSAB <- SST-SSA-SSB-SSE
  df.e <- (3*3)*(4-1)
  MSE <- SSE/df.e
  return(list("beta"=beta, "MSE"=MSE, "SSAB"=SSAB, "SSA"=SSA, "SSB"=SSB, "SSE"=SSE, "SSTO"=SST))
}
```

Testing our implementation vs. *aov* and *lm*.

```
fit_lm <-lm(data = data1914,V1~V2+V3+V2*V3) #fit linear regression
fit_aov <- summary(aov(data = data1914,V1~V2+V3+V2*V3)) #fit a anova model
fit_imp <-ANOVA33(data1914) #our fit
```

Our coefficients should match.

```
coe <- cbind(fit_imp$beta, fit_lm$coefficients)
colnames(coe) = c("Our Function","lm")
coe
```

```
##             Our Function    lm
## (Intercept)        2.475 2.475
## V22                2.975 2.975
## V23                3.500 3.500
## V32                2.125 2.125
## V33                2.100 2.100
## V22:V32            1.350 1.350
## V23:V32            2.175 2.175
## V22:V33            1.575 1.575
```

```
## V23:V33            5.175 5.175
```

Our sum of squares should match.

```
ss <- cbind(c(fit_imp$SSA, fit_imp$SSB, fit_imp$SSAB, fit_imp$SSE), fit_aov[[1]][["Sum Sq"]])
rownames(ss) = c("SSA","SSB","SSAB","SSE")
colnames(ss) = c("Our Function","aov")
ss
```

```
##      Our Function     aov
## SSA       220.020 220.020
## SSB       123.660 123.660
## SSAB       29.425  29.425
## SSE         1.625   1.625
```

And the MSE's should match.

```
ms <- cbind(fit_imp$MSE, fit_aov[[1]][["Mean Sq"]][4])
rownames(ms) = c("MSE")
colnames(ms) = c("Our Function","aov")
ms
```

```
##     Our Function        aov
## MSE   0.06018519 0.06018519
```

b)

Implement the estimation of contrasts, and of their standard errors.

```
contrast.95CI <- function(data,C,n = 36){
  beta.n <- n/9
  vector.matrix <- t(matrix(c(1,0,0,0,0,0,0,0,0,
                              1,1,0,0,0,0,0,0,0,
                              1,0,1,0,0,0,0,0,0,
                              1,0,0,1,0,0,0,0,0,
                              1,1,0,1,0,1,0,0,0,
                              1,0,1,1,0,0,1,0,0,
                              1,0,0,0,1,0,0,0,0,
                              1,1,0,0,1,0,0,1,0,
                              1,0,1,0,1,0,0,0,1),nrow = 9))
  #the index(i,j) for each cell
  cell.index <- matrix(c(1,1,1,2,2,2,3,3,3,1,2,3,1,2,3,1,2,3),ncol =2)
  #levels of value for V2 and V3
  level.v2 <- unique(data$V2)
  level.v3 <- unique(data$V3)
  design.matrix <- NULL
  mu.cell <- NULL
  for(i in 1:9){
    #data for each cell
    sub.data <- data%>%filter(V2 == cell.index[i,2],V3 == cell.index[i,1])
    #mu for each cell
    mu.cell[i] <- mean(sub.data$V1)
    #design matrix for each cell
    temp.matrix <- t(matrix(rep(vector.matrix[i,],nrow(sub.data)),ncol = 4))
    design.matrix <- rbind(design.matrix,temp.matrix)
  }
  y <- as.matrix(data$V1)
  beta <- solve(t(design.matrix)%*%design.matrix)%*%t(design.matrix)%*%y
  L <- C%*%beta
  e <- y-design.matrix%*%beta
  SSE <- sum(e^2)
  df.e <- nrow(data)-9
  MSE <- SSE/df.e
  se <- sqrt(MSE*C%*%solve(t(design.matrix)%*%design.matrix)%*%t(C))
```

```
  return(list("L"= L,"se"=se))
}
```

Now lets test it with $L = \mu_{12} - \mu_{13}$.

```
fit33 <- ANOVA33(data1914)
L = fit33$beta[4]-fit33$beta[5]
#because a=1, the contribution of alpha and interactions are 0
C <- matrix(c(0,0,0,1,-1,0,0,0,0),nrow = 1)
contrast <- contrast.95CI(data1914, C, fit33$MSE)
```

The L values match:

```
ls <- cbind(contrast$L, L)
rownames(ls) = c("L")
colnames(ls) = c("Our Function","Hand Calculated")
ls
```

```
##   Our Function Hand Calculated
## L        0.025           0.025
```

Not showing testing for SE as it requires other packages

Problem #3

Consider the dataset from KNNL problem 19.14

a)

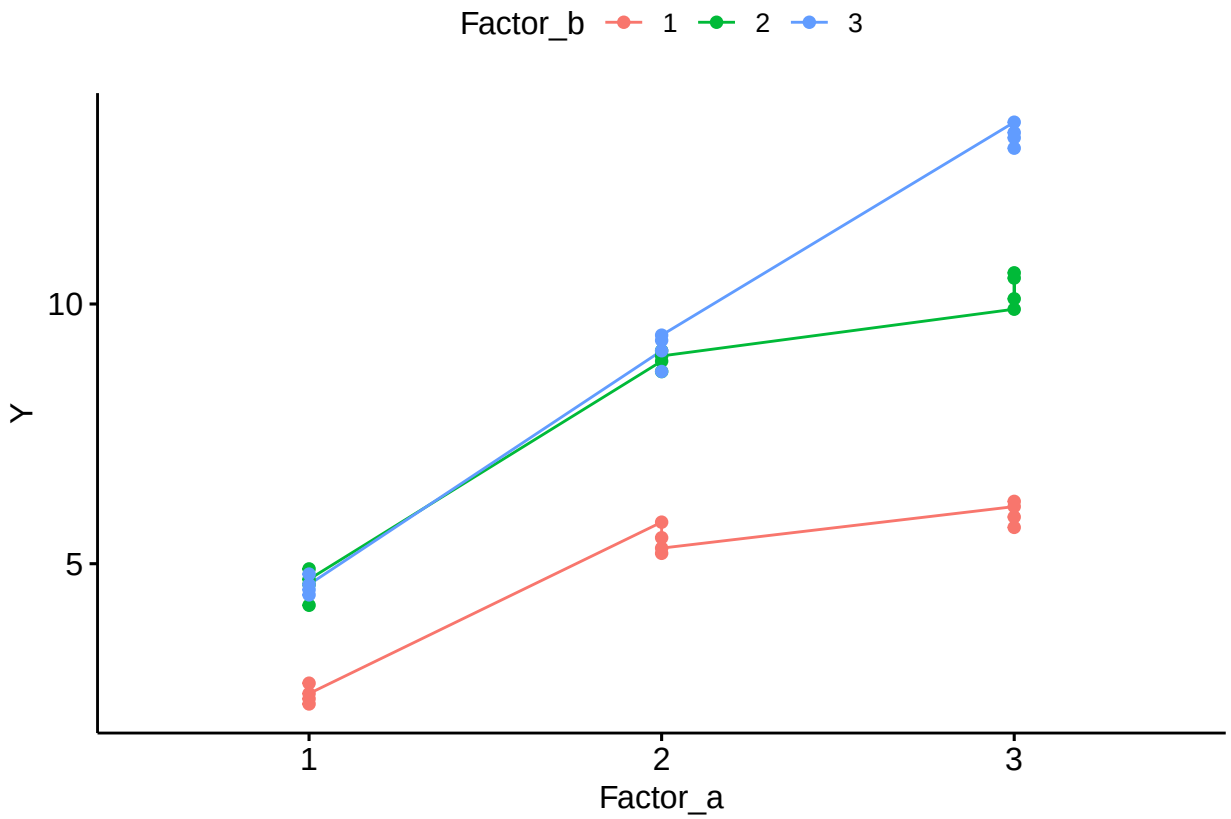
Briefly explain how to assign volunteers to treatments using an appropriate randomization, and why randomization is important.

Randomization is important to investigate the treatment effect, as it eliminates selection bias. Any differences we see after randomization are **not** from a systematic bias. After randomization, **each participant's experimental condition is independent of each other.** True randomness is hard to come by and ideally the randomization scheme produces balanced groups with similar sample sizes. A good approach is to use a randomization tool to randomize a sequence of assignments thus ensuring balanced groups and randomness.

b)

Visualize the dataset with a treatment means plot with Factor A as the x axis and Factor B as the line/point color. Based on the plot, comment on whether both factors and the interaction are likely to be present.

```
data1914 <- data1914%>%mutate(Y = V1,Factor_a=V2,Factor_b=V3)
ggline(data1914, x = "Factor_a", y = "Y", color = "Factor_b")
```



Since the lines are not parallel, an interaction is likely to be present.

c)

Using your implementation in Question 2, test for the presence of interactions, at the confidence level of 95%. Please state the hypotheses and interpret your results.

$$H_0 : \text{All } (\alpha\beta)_{ij} = 0$$
$$H_a : \text{Not all } (\alpha\beta)_{ij} = 0$$

$$\frac{\frac{SSAB}{(a-1)(b-1)}}{\frac{SSE}{(ab)(n-1)}} \sim F_{(a-1)(b-1),ab(n-1)}$$

```
fit33 <- ANOVA33(data1914)
df_num = (3-1)*(3-1)
df_denom <- (3*3)*(4-1)
F_score <- (fit33$SSAB/df_num)/(fit33$SSE/df_denom)
qf <- qf(df1 = df_num,df2 = df_denom,p = 0.95)
print(paste("F_score:",F_score,"larger than",qf,"reject null hypothesis" ))
```

```
## [1] "F_score: 122.226923076923 larger than 2.72776530603399 reject null hypothesis"
```

This F test compares the MSAB to the MSE, and finds that the MSAB. As the F_score of 122.2269231 is larger than the critical F value of 2.7277653, we can reject H_0 with 95% confidence. This means there is evidence that the one or more $(\alpha\beta)_{ij}$ is significantly difference from zero, and there is likely interaction present somewhere between levels of factors A and B.

d)

Using your implementation in Question 2, estimate μ_{23} with a 95% confidence interval and interpret the results.

From Question 1j), we know that μ_{23} has vector coefficients of C = (1,1,0,0,1,0,0,1,0).

```
#parameter vector
C <- matrix(c(1,1,0,0,1,0,0,1,0),nrow = 1)
fit <- contrast.95CI(data1914,C) #point estimate and SE
df.e <- (4-1)*(3*3) # ab(n-1)
t <- qt(0.975,df.e)
CI<-round(c(fit$L-t*fit$se,fit$L+t*fit$se),digits = 2) #get 95% CI
```

μ_{23} is the estimated mean response when factor A is 2 and factor B is 3. The 95% lower bound for μ_{23} is 8.87. The 95% upper bound for μ_{23} is 9.38.

e)

Using your implementation in Question 2, estimate $L = \mu_{12} - \mu_{13}$ with a 95% confidence interval and interpret the results.

From Question 1k), we know that $\mu_{12} - \mu_{13}$ has vector coefficients of C = (0,0,0,1,-1,0,0,0,0)

```
#parameter vector
C <- matrix(c(0,0,0,1,-1,0,0,0,0),nrow = 1)
fit <- contrast.95CI(data1914,C)#point estimate and SE
df.e <- (4-1)*(3*3) # ab(n-1)
t <- qt(0.975,df.e)
CI<-round(c(fit$L-t*fit$se,fit$L+t*fit$se), digits = 4)#get 95% CI
```

$\mu_{12} - \mu_{13}$ is the estimated difference between the mean responses between factor B=2 and B=3, while factor A is 2. The 95% lower bound for $\mu_{12} - \mu_{13}$ is -0.3309. The 95% upper bound for $\mu_{12} - \mu_{13}$ is 0.3809.

Problem #4

Consider the dataset from KNNL problem 21.5.

a)

Explain why a randomized complete block design is useful in this problem.

Randomized complete block design makes the treatment of each unit independent from which block they are. In this problem, the variable "graduate time" could potentially affect training results, and our sample size is relatively small. Use of blocking allows us to control for the effects of the blocking variable.

b)

Visualize the dataset with a treatment means plot, and conduct the Tukey test for additivity. State the null and the alternative hypotheses, and the conclusion. You may use regression libraries like lm or anova for this question.

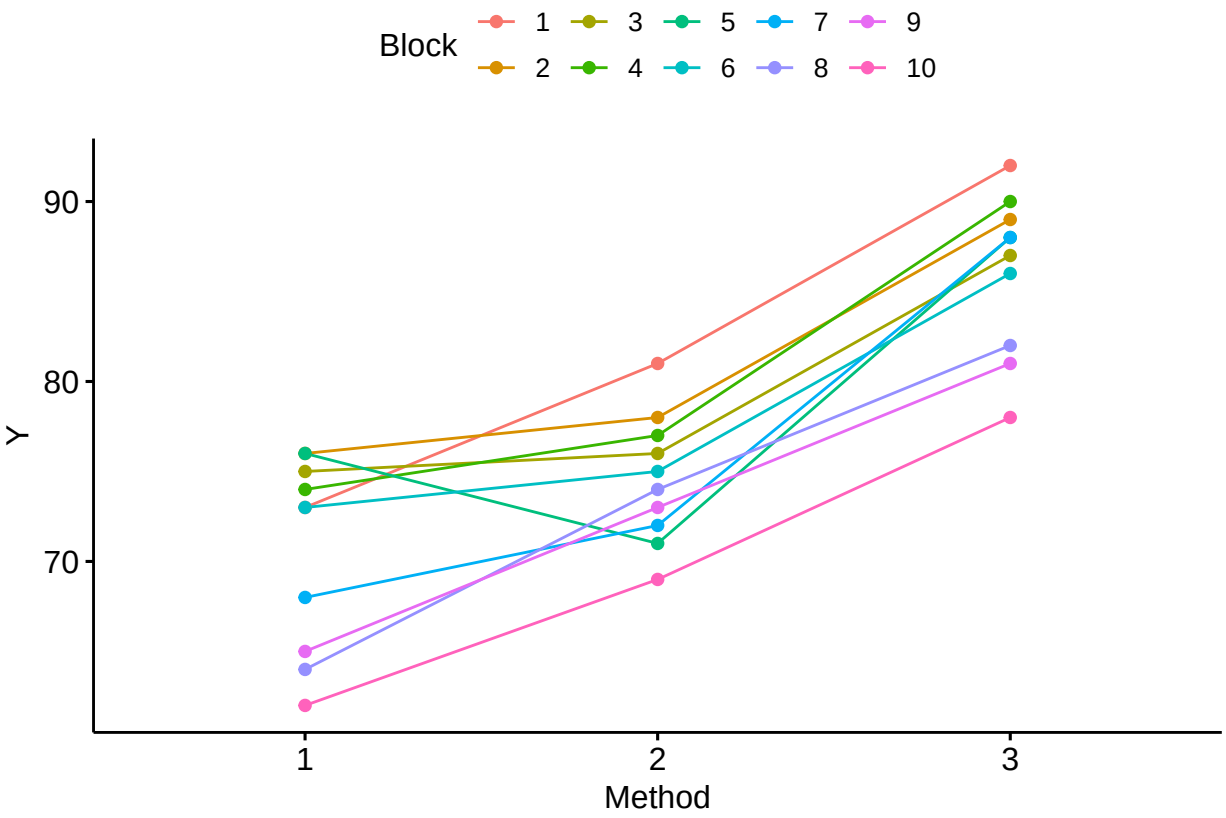
The Model

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}, \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

where $\alpha_1 = \beta_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$

$H_0 : (\alpha\beta)_{ij} = 0$ for all i, j, $H_a : (\alpha\beta)_{ij} \neq 0$ for some i, j

```
data2105 <- fread("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapte
data2105$V2 <- as.factor(data2105$V2)
data2105$V3 <- as.factor(data2105$V3)
data2105 <- data2105%>%mutate(Y = V1,Method = V3,Block = V2)
ggline(data2105, x = "Method", y = "Y", color = "Block")
```



Although each block has **roughly** the same linear upwards trend, their average overall Y responses vary.

Regression approach:

```
fitb1 <- lm(V1~V2,data = data2105)
fitb2 <- lm(V1~V3,data = data2105)
# basically fit each model as a one factor linear regression model,
data2105$alphabeta <- c(0,0,0,as.vector(t(matrix(rep(fitb1$coefficients[2:10],3),ncol = 3)))) *
rep(c(0,fitb2$coefficients[2:3]),10)
fitb3 <- lm(V1~V2+V3+alphabeta,data = data2105)
summary(fitb3)
```

```
##
## Call:
## lm(formula = V1 ~ V2 + V3 + alphabeta, data = data2105)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8414 -1.0822 -0.5156  1.4596  4.1458
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  75.582858   1.731496  43.652  < 2e-16 ***
## V22          -1.016910   2.101231  -0.484 0.634590
## V23          -2.711760   2.122798  -1.277 0.218616
## V24          -1.694850   2.107528  -0.804 0.432391
## V25          -3.728670   2.144920  -1.738 0.100222
## V26          -4.067640   2.153782  -1.889 0.076130 .
## V27          -6.101459   2.221918  -2.746 0.013782 *
## V28          -8.813219   2.349458  -3.751 0.001591 **
## V29          -9.152189   2.368066  -3.865 0.001243 **
## V210        -12.541889   2.582208  -4.857 0.000148 ***
## V32           3.949010   1.206492   3.273 0.004483 **
## V33          15.302414   1.831761   8.354 2.01e-07 ***
## alphabeta    -0.002602   0.018784  -0.138 0.891474
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.569 on 17 degrees of freedom
## Multiple R-squared:  0.939,  Adjusted R-squared:  0.896
## F-statistic: 21.82 on 12 and 17 DF,  p-value: 5.739e-08
```

ANOVA approach:

```
(anova <- anova(fitb3))
```

```
## Analysis of Variance Table
##
## Response: V1
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## V2         9  433.37   48.15  7.2953 0.0002466 ***
## V3         2 1295.00  647.50 98.1002 4.614e-10 ***
## alphabeta  1    0.13    0.13  0.0192 0.8914739
## Residuals 17  112.21    6.60
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
F_score <- anova[3,3]/anova[4,3]
F_cutoff <- qf(p = 0.95,df1 = 1,df2 = 30-10-3)
```

F statistic: 0.0191818 Critical F value: 4.4513218 F statistic > Critical F value: FALSE

Note that the two approaches are equivalent, as the $T_a \times T_b = F_{a,b}$ distribution : $(-0.138)^2 = 0.019$

Fail to reject null hypothesis. There is no evidence *against* an additive effect.

c)

Use a standard implementation of linear models in R (or any other language) to fit the additive model to the data. Interpret your results.

Fit an additive model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

where $\alpha_1 = \beta_1 = 0$

```
fit4c <- lm(data2105,formula = Y~Block+Method)
print(summary(fit4c))
```

```
##
## Call:
## lm(formula = Y ~ Block + Method, data = data2105)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.833 -1.125 -0.500  1.500  4.167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   75.500      1.580  47.786  < 2e-16 ***
## Block2        -1.000      2.040  -0.490 0.629872
## Block3        -2.667      2.040  -1.307 0.207544
## Block4        -1.667      2.040  -0.817 0.424554
## Block5        -3.667      2.040  -1.798 0.089033 .
## Block6        -4.000      2.040  -1.961 0.065533 .
## Block7        -6.000      2.040  -2.942 0.008725 **
## Block8        -8.667      2.040  -4.249 0.000483 ***
## Block9        -9.000      2.040  -4.412 0.000336 ***
## Block10      -12.333      2.040  -6.047 1.02e-05 ***
## Method2        4.000      1.117   3.580 0.002139 **
```

```
## Method3       15.500      1.117  13.874 4.72e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.498 on 18 degrees of freedom
## Multiple R-squared:  0.939,  Adjusted R-squared:  0.9017
## F-statistic: 25.18 on 11 and 18 DF,  p-value: 1.107e-08
```

```
print(anova(fit4c))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## Block      9  433.37   48.15   7.7157 0.0001316 ***
## Method     2 1295.00  647.50 103.7537 1.315e-10 ***
## Residuals 18  112.33    6.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model has a large R^2 and a significant F stat (25.18 on 11 and 18, $p < 0.001$) indicating it explains much of the variability in the data. Looking at the anova output, we can see that method is significant.

d)

Use a standard implementation of linear models in R (or any other language) to derive a confidence interval for the difference between the training methods 1 and 2. Interpret your results.

```
confint(fit4c,"Method2",level = 0.95)
```

```
##             2.5 %   97.5 %
## Method2 1.652838 6.347162
```

"Method2" is the second level of variable Method, as compared to the baseline (Method1). The 95% CI for Training method 2 is between 1.65 to 6.34.

e)

Repeat (d), while ignoring the blocking (i.e., treat the data as if it came from a completely randomized experiment). Comment on the difference in the width of the confidence intervals.

First refit the model with just method, then generate the intervals.

```
fit4e <- lm(Y~Method,data = data2105)
summary(fit4e)
```

```
##
## Call:
## lm(formula = Y ~ Method, data = data2105)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.600 -3.350  1.150  3.275  6.400
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   70.600      1.422  49.660  < 2e-16 ***
## Method2        4.000      2.011   1.990   0.0569 .
## Method3       15.500      2.011   7.709 2.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.496 on 27 degrees of freedom
## Multiple R-squared:  0.7035, Adjusted R-squared:  0.6816
```

```
## F-statistic: 32.04 on 2 and 27 DF,  p-value: 7.441e-08
```

```
round(confint(fit4e,"Method2",level = 0.95), digits=2)
```

```
##         2.5 % 97.5 %
## Method2 -0.13   8.13
```

The confidence interval is wider compare to part d). Removing blocking from the model increases the standard error, and thus the uncertainty - meaning the 95 CI will be larger.

f)

Repeat (d), while ignoring training method 3 (i.e., treat the data as if the experiment did not contain the third training). Comment on the difference in the width of the confidence intervals.

First refit the model with out observations where method=3, then generate the intervals.

```
data2105_f <- data2105%>%filter(Method!=3)
fit4f <- lm(Y~Method+Block,data = data2105_f)
summary(fit4f)
```

```
##
## Call:
## lm(formula = Y ~ Method + Block, data = data2105_f)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -4.5   -1.5    0.0    1.5    4.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.500e+01  2.319e+00  32.341 1.27e-10 ***
## Method2      4.000e+00  1.398e+00   2.860  0.01877 *
## Block2      -2.150e-14  3.127e+00   0.000  1.00000
## Block3      -1.500e+00  3.127e+00  -0.480  0.64288
## Block4      -1.500e+00  3.127e+00  -0.480  0.64288
## Block5      -3.500e+00  3.127e+00  -1.119  0.29199
## Block6      -3.000e+00  3.127e+00  -0.959  0.36242
## Block7      -7.000e+00  3.127e+00  -2.239  0.05196 .
## Block8      -8.000e+00  3.127e+00  -2.558  0.03077 *
## Block9      -8.000e+00  3.127e+00  -2.558  0.03077 *
## Block10     -1.150e+01  3.127e+00  -3.678  0.00509 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.127 on 9 degrees of freedom
## Multiple R-squared:  0.8048, Adjusted R-squared:  0.5879
## F-statistic:  3.71 on 10 and 9 DF,  p-value: 0.0306
```

```
round(confint(fit4f,"Method2",level = 0.95), digits=2)
```

```
##         2.5 % 97.5 %
## Method2  0.84   7.16
```

Again, the confidence interval is wider compare to part d). By removing data, we decrease n and increase the estimated standard error.