HW 5

Andrew Jowe

1

a. $R^2\{AB|(A+B)\} = 0.0052$

b. The value in a does not suggest interaction between factors will play an important role because this value is small (close to 0).

c. H_0 : All $(\gamma\delta)_{ij} = 0$

   H_A : At least one $(\gamma\delta)_{ij} \neq 0$

   $F_s = 2.2093$

   $p = 0.1379$

   Since $p > \alpha$, we fail to reject H_0. Therefore, there is no interaction effect between our two factors.

d. The p-value represents the type I error, which is the probability that we reject H_0 when in reality H_0 is true. In other words, it's the probability that we claim that there is an interaction effect when in reality there is no interaction effect.

2

a. $\bar{Y}_{12.}$ is the sample mean for all observations within low dose A and medium dose B.
b. $\bar{Y}_{12.}$ is the sample mean for all observations within low dose A.
c. $R^2\{A|\cdot\} = 0.5871$
d. $R^2\{B|\cdot\} = 0.33$
e. Factor A seems to be more important because the R^2 value for it is larger than factor B.

3

a. H_0 : All $(\gamma\delta)_{ij} = 0$ H_A : At least one $(\gamma\delta)_{ij} \neq 0$ $F_s = 121.8313$ $p = 0$
b. Since $p \leq \alpha$, we reject H_0. Therefore, there is an interaction effect between our two factors.
c. The type I error is the probability that we reject H_0 when in reality H_0 is true. In other words, it's the probability that we claim that there is an interaction effect when in reality there is no interaction effect.
d. No, we would not test for individual effects of A and B, because we cannot simplify our model due to rejecting H_0. This is only needed if we accept H_0 to see if we can simplify the model any further.

4

a. $\bar{Y}_{11.} - \bar{Y}_{12.}$ represents the difference between the sample mean of all observations within female no smoking and the sample mean of all observations within female yes smoking.

b. $\bar{Y}_{11.} - \bar{Y}_{21.}$ represents the difference between the sample mean of all observations within female no smoking and the sample mean of all observations within male no smoking.
c. We want smallest type I error, so $\alpha = 0.001$. Since $p \geq \alpha$, we fail to reject H_0. Therefore, there is no interaction effect between the two factors.
d. $R^2\{AB|(A{+}B)\} = 0.0366$ Our value seems reasonable because we concluded that there is no interaction effect which is typical for small values of R^2 (close to 0).

5

a. H_0 : All $\gamma_i = 0$ H_A : At least one $\gamma_i \neq 0$ $F_s = 16.0916$ $p = 10^{-4}$ Since $p \leq \alpha$, we reject H_0. Therefore, there is a factor A effect.

b. H_0 : All $\gamma_i = 0$ H_A : At least one $\gamma_i \neq 0$ $F_s = 42.7493$ $p = 0$ Since $p \leq \alpha$, we reject H_0. Therefore, there is a factor B effect.

c. $R^2\{(A + B)|B\} = 0.0904$ $R^2\{(A + B)|A\} = 0.2088$

d. I would recommend to use the model that includes factor A and factor B effect, but excludes the interaction effect. For the interaction effect, our p-value and R^2 both suggest that there is no effect. For factor A and factor B effect, our p-value suggests that there is an effect while our R^2 does not suggest an effect due to the low value (close to 0). However, the F-statistic test is an actual test while the R^2 value only suggests what we might want to do, so the F-statistic test takes precedence. This reason combined with wanting to be conservative, we should include both factor A and factor B effect. The model is as follows: $Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + \epsilon_{ijk}$

6

a. False, while it may suggest which model we might want to use, using the F-statistic test to determine that is more sound and thus takes precedence.
b. True, the factor A effect is what causes at least one pair-wise comparison of factor A means to have a significant difference.
c. True, we can see this in the equation as follows: $R^2 = \frac{SSE_F - SSE_R}{SSE_F}$ In the equation, we are assuming that the full model always has more unexplained variance (SSE) than the reduced model. Hence, the numerator is always less than or equal to the denominator.
d. False, as total population variance is unknown, we use SSE to estimate it, i.e. $\sigma_\epsilon^2 = SSE$. However, SSE differs between models, so our total variance also differs. This means that the variance for each observation differs depending on the model we choose.

I

a. $n_T = 60$, $a = 2$, $b = 3$
b. $SSE\{AB\} = 1.1586 \times 10^4$ $SSE\{A + B\} = 1.2764133 \times 10^4$
c. The table of all $\bar{Y}_{ij}$ is as follows:

A	B	$\bar{Y}_{ij.}$
High	Beef	100.0
Low	Beef	79.2
High	Cereal	85.9
Low	Cereal	83.9

A	B	$\bar{Y}_{ij.}$
High	Pork	99.5
Low	Pork	78.7

d. $F_s = 2.7455$ $p = 0.0732$
e. Since $p > \alpha$, we fail to reject H_0. Therefore, there is no interaction effect.

II

a. $R^2\{(A+B)|B\} = 0.1989$ $R^2\{(A+B)|A\} = 0.0205$
b. $\bar{Y}_{H.} - \bar{Y}_{L.} = 14.5333$ $\bar{Y}_{.B} - \bar{Y}_{.P} = 0.5$
c. $F_s = 13.9001$ $p = 5 \times 10^{-4}$ Since $p \leq \alpha$, we reject H_0. Therefore, there is a factor A effect.
d. $F_s = 0.5847$ $p = 0.5606$ Since $p > \alpha$, we fail to reject H_0. Therefore, there is no factor B effect.
e. The final model we suggest using is the one using factor A effect only because from our F-statistic tests, we conclude that there is only factor A effect, no factor B effect, and no interaction effect. The model is as follows: $Y_{ij} = \mu_{.} + \gamma_i + \epsilon_{ij}$

III

a. $n_T = 24$, $a = 2$, $b = 3$
b. $SSE\{AB\} = 1550.25$ $SSE\{A+B\} = 2.1715583 \times 10^4$ $SSE\{A\} = 5.8127583 \times 10^4$ $SSE\{B\} = 6.1162625 \times 10^4$ $SSE\{Empty\} = 9.7574625 \times 10^4$
c. $R^2\{AB|(A+B)\} = 0.9286$
d. The both, exp group had the lowest average time to complete the project.

A	B	$\bar{Y}_{ij.}$
Both	Exp	38.75
Small	Exp	60.50
Both	Med	44.75
Small	Med	106.50
Both	New	62.25
Small	New	222.00

IV

a. $F_s = 117.0701$ $p = 0$ Since $p \leq \alpha$, we reject H_0. Therefore, there is an interaction effect and we should use the full model as follows: $Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \epsilon_{ijk}$
b. $R^2\{A|\cdot\} = 0.4043$ $R^2\{B|\cdot\} = 0.3732$
c. The values in b suggests that the main effects should be included because these are large.
d. I would suggest a programmer in the both, exp treatment group to complete a project as soon as possible because it has the lowest average completion time and our results show that both of these factors from the treatment do affect the completion time.

Appendix

Libraries

```
rat <- read.csv("rat.csv")
prog <- read.csv("Prog.csv")
```

Functions

```
Partial.R2 = function(small.model,big.model){
  SSE1 = sum(small.model$residuals^2)
  SSE2 = sum(big.model$residuals^2)
  PR2 = (SSE1 - SSE2)/SSE1
  return(PR2)
}
```

1

```
# a
SSE <- c("AB" = 676.80, "A+B" = 680.31)
R_squared <- c("AB|(A+B)" = (SSE[["A+B"]] - SSE[["AB"]]) / SSE[["A+B"]])

# c
n_T <- 430
a <- 2
b <- 2

SSE_F <- SSE["AB"]
SSE_R <- SSE["A+B"]
df_SSE_F <- n_T - a * b
df_SSE_R <- n_T - a - b + 1

MSE_F <- SSE_F / df_SSE_F
F_s <- (SSE_R - SSE_F) / (df_SSE_R - df_SSE_F) / MSE_F
df_num <- df_SSE_R - df_SSE_F
df_den <- df_SSE_F

the.p.value <- pf(F_s, df_num, df_den, lower.tail = FALSE)
```

2

```
SSE <- c("AB" = 1.63, "A+B" = 31.05, "A" = 154.71,
         "B" = 251.07, "Null" = 374.73)
R_squared <- c(
  "A|Null" = (SSE[["Null"]] - SSE[["A"]]) / SSE[["Null"]],
  "B|Null" = (SSE[["Null"]] - SSE[["B"]]) / SSE[["Null"]]
)
```

3

```r
# using SSE from 2
n_T <- 36
a <- 3
b <- 3

SSE_F <- SSE["AB"]
SSE_R <- SSE["A+B"]
df_SSE_F <- n_T - a * b
df_SSE_R <- n_T - a - b + 1

MSE_F <- SSE_F / df_SSE_F
F_s <- (SSE_R - SSE_F) / (df_SSE_R - df_SSE_F) / MSE_F
df_num <- df_SSE_R - df_SSE_F
df_den <- df_SSE_F

the.p.value <- pf(F_s, df_num, df_den, lower.tail = FALSE)
```

4

```r
SSE <- c("AB" =  682.43, "A+B" = 708.34, "A" = 895.26, "B" = 778.70)
R_squared <- c("AB|(A+B)" = (SSE[["A+B"]] - SSE[["AB"]]) / SSE[["A+B"]])
```

5-a

```r
# using SSE from 4
n_T <- 165
a <- 2
b <- 2

SSE_F <- SSE["A+B"]
SSE_R <- SSE["B"]
df_SSE_F <- n_T - a - b + 1
df_SSE_R <- n_T - b

MSE_F <- SSE_F / df_SSE_F
F_s <- (SSE_R - SSE_F) / (df_SSE_R - df_SSE_F) / MSE_F
df_num <- df_SSE_R - df_SSE_F
df_den <- df_SSE_F

the.p.value <- pf(F_s, df_num, df_den, lower.tail = FALSE)
```

5-b

```
# using SSE from 4
n_T <- 165
a <- 2
b <- 2

SSE_F <- SSE["A+B"]
SSE_R <- SSE["A"]
df_SSE_F <- n_T - a - b + 1
df_SSE_R <- n_T - a

MSE_F <- SSE_F / df_SSE_F
F_s <- (SSE_R - SSE_F) / (df_SSE_R - df_SSE_F) / MSE_F
df_num <- df_SSE_R - df_SSE_F
df_den <- df_SSE_F

the.p.value <- pf(F_s, df_num, df_den, lower.tail = FALSE)
```

5-c

```
# using SSE from 4
R_squared <- c(
  "(A+B)|B" = (SSE[["B"]] - SSE[["A+B"]]) / SSE[["B"]],
  "(A+B)|A" = (SSE[["A"]] - SSE[["A+B"]]) / SSE[["A"]]
)
```

I

```
# a
n_T <- nrow(rat)
a <- length(unique(rat$Amount))
b <- length(unique(rat$Type))

# b
the.data <- rat
names(the.data) = c("Y","A","B")

AB = lm(Y ~ A*B,the.data)
A.B = lm(Y ~ A + B,the.data)
A = lm(Y ~ A,the.data)
B = lm(Y ~ B,the.data)
N = lm(Y ~ 1, the.data)

all.models = list(AB,A.B,A,B,N)
SSE = t(as.matrix(sapply(all.models,function(M) sum(M$residuals^2))))
colnames(SSE) = c("AB","(A+B)","A","B","Empty/Null")
rownames(SSE) = "SSE"

# c
means <- aggregate(Y ~ A + B, data = the.data, FUN = mean)
```

```r
colnames(means)[which(names(means) == "Y")] <- "$\\bar Y_{ij.}$"

# d
the.anova.result <- anova(A.B, AB)
the.f.statistic <- the.anova.result$"F"[2]
the.p.value <- the.anova.result$"Pr(>F)"[2]
```

II

```r
# a
R_squared <- c("(A+B)|B" = Partial.R2(B, A.B), "(A+B)|A" = Partial.R2(A, A.B))

# b
means_A <- aggregate(Y ~ A, data = the.data, FUN = mean)
diff.means_A <- means_A$Y[means_A$A == "High"] - means_A$Y[means_A$A == "Low"]

means_B <- aggregate(Y ~ B, data = the.data, FUN = mean)
diff.means_B <- means_B$Y[means_B$B == "Beef"] - means_B$Y[means_B$B == "Pork"]

# c
the.anova.result_A <- anova(B, A.B)
the.f.statistic_A <- the.anova.result_A$"F"[2]
the.p.value_A <- the.anova.result_A$"Pr(>F)"[2]

# d
the.anova.result_B <- anova(A, A.B)
the.f.statistic_B <- the.anova.result_B$"F"[2]
the.p.value_B <- the.anova.result_B$"Pr(>F)"[2]
```

III

```r
# a
n_T <- nrow(prog)
a <- length(unique(prog$type))
b <- length(unique(prog$years))

# b
the.data <- prog
names(the.data) = c("Y","A","B")

AB = lm(Y ~ A*B,the.data)
A.B = lm(Y ~ A + B,the.data)
A = lm(Y ~ A,the.data)
B = lm(Y ~ B,the.data)
N = lm(Y ~ 1, the.data)

all.models = list(AB,A.B,A,B,N)
SSE = t(as.matrix(sapply(all.models,function(M) sum(M$residuals^2))))
colnames(SSE) = c("AB","(A+B)","A","B","Empty/Null")
```

```
rownames(SSE) = "SSE"

# c
R_squared <- c("AB|(A+B)" = Partial.R2(A.B, AB))

# d
means <- aggregate(Y ~ A + B, data = the.data, FUN = mean)
colnames(means)[which(names(means) == "Y")] <- "$\\bar Y_{ij.}$"
```

IV

```
# a
the.anova.result <- anova(A.B, AB)
the.f.statistic <- the.anova.result$"F"[2]
the.p.value <- the.anova.result$"Pr(>F)"[2]

# b
R_squared[["A|N"]] <- Partial.R2(N, A)
R_squared[["B|N"]] <- Partial.R2(N, B)
```