MET CS 555 Assignment 5 – 20 points Fall 2022

**SUBMISSION REQUIREMENTS: Please submit a single document (word or PDF) for submission.  Your submission should contain a summary of your results (and answers to questions asked on the homework) as well as your R code used to generate your results (please append to the end of your submission). Please use R for the calculations whenever possible. You will lose points if you are not utilizing R.**

**The data in this document is from 3 groups of students (math, chemistry, and physics) on an IQ related test.** **Save the data to excel and read the data into R. Use this data to address the following questions:**

1. How many students are in each group? Summarize the data relating to both test score and age by the student group (separately). Use appropriate numerical and/or graphical summaries. – 3 points

|  |  |
| --- | --- |
| Chemistry students | 15 |
| Math students | 15 |
| Physics students | 15 |

**Chemistry students**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
| 40 | 44 | 46 | 46.27 | 48 | 52 |

Chart, scatter chart

Description automatically generated

**Math students**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
| 24 | 36 | 38 | 37.6 | 40.5 | 45 |

Chart, scatter chart

Description automatically generated

**Physics students**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
| 25 | 31.5 | 34 | 34.13 | 37.5 | 42 |

Chart, scatter chart

Description automatically generated

1. Do the test scores vary by student group? Perform a one-way ANOVA using the aov or Anova function in R to assess. Use a significance level of α=0.05. Summarize the results using the 5-step procedure. If the results of the overall model are significant, perform the appropriate pairwise comparisons using Tukey’s procedure to adjust for multiple comparisons and summarize these results. – 7 points

**🡪 One-way Anova**

**Hypothesis**

H0 ∶ μ1 = μ2 = μ3

H1 : μ1 ≠ μ2 ≠ μ3

α = 0.05

**2.2 - Select test statistic**

F= MSB/MSW

k-1 = 2

n-k = 45 - 3 = 42 deg of freedom

**2.3 - State decision rule**

Decision Rule: Reject H0 if F ≥ 3.219942

Otherwise, do not reject H0

**2.4 - Compute test statistic**

F = 26.57

**2.5 – Conclusion**

We have enough evidence to reject null hypothesis given that F value (26.57) > 3.219942

**🡪 Pairwise comparison**

**Text

Description automatically generated**

**Interpretation**

The mean of the Math students is 8.6667 less compared to the mean of the Chemistry students. Additionally, the mean of the Physics students if 12.1333 less than the mean of the Chemistry students and 3.46667 less than the mean of the Math students.

1. Create an appropriate number of dummy variables for student group and re-run the one-way ANOVA using the lm function with the newly created dummy variables. Set chemistry students as the reference group. Confirm if the results are the same as in Q2. What is the interpretation of the beta estimates from the regression model? – 4 points

Text

Description automatically generated

**Interpretation**

The mean for the Chemistry students is 46.267, while the mean of the Math and Physics students is 8.667 and 12.133 less respectively than the mean of the Chemistry students. These results are exactly the same than the results obtained in Q2.

1. Re-do the one-way ANOVA adjusting for age. Focus on the output relating to the comparisons of test score by student type. Explain how this analysis differs from the analysis in step 2 above (not the results but how does this analysis differ in terms of the question that it is trying to answer). Did you obtain different results? Briefly summarize (no need to go through the 5–step procedure here). Present the least square means and interpret these. – 6 points

~~Text

Description automatically generated~~

**Interpretation**

After re-doing the analysis adjusting for age, the group still has a significant role in estimating IQ. However, the difference in means between each group is very different compared to previous analysis, which suggests that adjusting for age had a relevant effect on the results.

Text

Description automatically generated

**Interpretation**

The least square means analysis shows that the difference in means between all groups is relatively small, given that the minimum mean is 38.6 (Chemistry students) and the maximum mean is 40.5 (Math students).

**R Code**

data <- read.csv("data.csv", header=TRUE)

attach(data)

#Problem 1

table(group)

chem <- data[data$group == "Chemistry student",]

math <- data[data$group == "Math student",]

phys <- data[data$group == "Physics student",]

summary(chem$iq)

summary(math$iq)

summary(phys$iq)

install.packages("ggplot2")

require(ggplot2)

ggplot(chem, aes(x=chem$age,y=chem$iq)) +

geom\_point(shape=1,color="red") + xlab("Age") + ylab("IQ score") + xlim(c(min(chem$age)-2,max(chem$age))) + ylim(c(min(chem$iq)-2,max(chem$iq))) +

ggtitle("Scatterplot of age and IQ score in Chemistry students") + theme\_bw(base\_size=14)

ggplot(math, aes(x=math$age,y=math$iq)) +

geom\_point(shape=1,color="red") + xlab("Age") + ylab("IQ score") + xlim(c(min(math$age)-2,max(math$age))) + ylim(c(min(chem$iq)-2,max(chem$iq))) +

ggtitle("Scatterplot of age and IQ score in Math students") + theme\_bw(base\_size=14)

ggplot(phys, aes(x=phys$age,y=phys$iq)) +

geom\_point(shape=1,color="red") + xlab("Age") + ylab("IQ score") + xlim(c(min(phys$age)-2,max(phys$age))) + ylim(c(min(phys$iq)-2,max(chem$iq))) +

ggtitle("Scatterplot of age and IQ score in Physics students") + theme\_bw(base\_size=14)

#Problem 2 -

#2.1 - Hypothesis

#H0 ∶ μ1 = μ2 = μ3

#H1 : μ1 ≠ μ2 ≠ μ3

#α = 0.05

#2.2 - Select test statistic

#F= MSB/MSW

#k-1 = 2

#n-k = 45 - 3 = 42 deg of freedom

#2.3 - State decision rule

qf(.95, df1=2, df2=42)

#Decision Rule: Reject H0 if F ≥ 3.219942

#Otherwise, do not reject H0

#2.4 - Compute test statistic

anova.model <- aov(data$iq~group, data=data)

summary(anova.model)

#2.5 - Conclusion

#We have enough evidence to reject null hypothesis given that F value (26.57) > 3.219942

#2.6 - Tukey

TukeyHSD(anova.model)

#Problem 3

data$g1 <- ifelse(data$group == "Math student",1,0)

data$g2 <- ifelse(data$group == "Physics student",1,0)

model <- lm(data$iq~data$g1+data$g2, data=data)

summary(model)

#Problem 4

data$group = factor(data$group)

library(car)

mdl = lm(iq ~ group + age,data=data)

summary(mdl)

library(emmeans)

emmeans(mdl, specs = "group")