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.

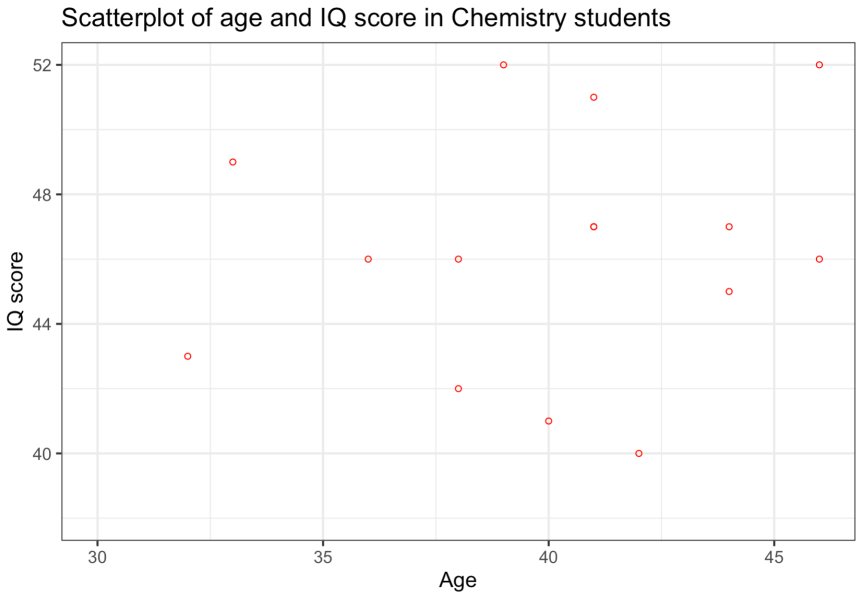
The data in this document is from 3 groups of students (math, chemistry, and physics) on an IQ related test. Save the data to excel and read the data into R. Use this data to address the following questions:

(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

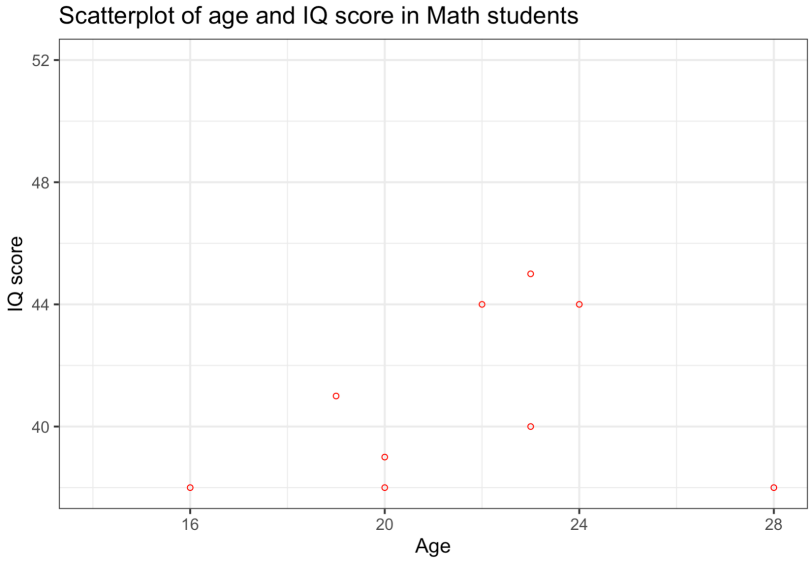
Chemistry students

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



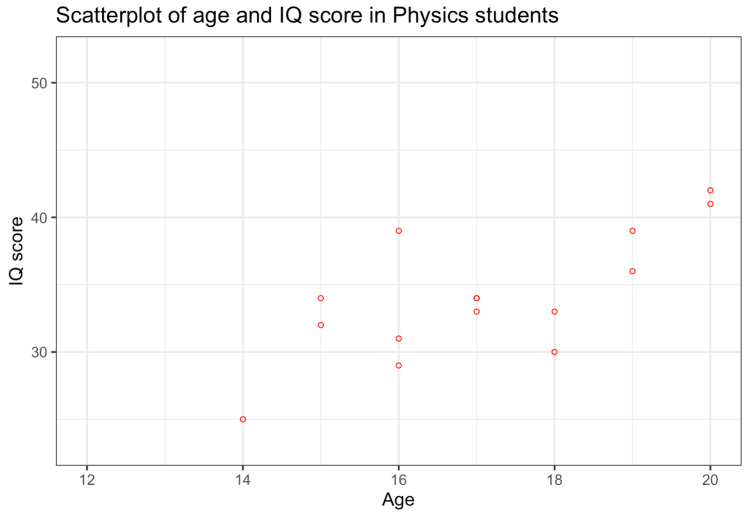
Math students

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



Physics students

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



(2) 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

```
> #2.6 - Tukey
> TukeyHSD(anova.model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = data$iq ~ group, data = data)

$group
                                      diff        lwr        upr     p adj
Math student-Chemistry student     -8.666667 -12.832756 -4.5005778 0.0000262
Physics student-Chemistry student -12.133333 -16.299422 -7.9672445 0.0000000
Physics student-Math student       -3.466667  -7.632756  0.6994222 0.1194835
```

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.

(3) 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

```
> summary(model)

Call:
lm(formula = data$iq ~ data$g1 + data$g2, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-13.6000  -2.1333  -0.1333   2.7333   7.8667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.267      1.213  38.157  < 2e-16 ***
data$g1       -8.667      1.715  -5.054 8.93e-06 ***
data$g2      -12.133      1.715  -7.076 1.13e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.696 on 42 degrees of freedom
Multiple R-squared:  0.5585,    Adjusted R-squared:  0.5375
F-statistic: 26.57 on 2 and 42 DF,  p-value: 3.496e-08
```

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.

(4) 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

```
> summary(mdl)

Call:
lm(formula = iq ~ group + age, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1032  -2.5127   0.2222   2.9920   6.6030

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          24.3263     8.6939   2.798   0.0078 **
groupMath student     1.9202     4.4606   0.430   0.6691
groupPhysics student  0.4249     5.1901   0.082   0.9352
age                   0.5476     0.2151   2.546   0.0148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.417 on 41 degrees of freedom
Multiple R-squared:  0.6188,    Adjusted R-squared:  0.5909
F-statistic: 22.18 on 3 and 41 DF,  p-value: 1.078e-08
```

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.

```
> emmeans(mdl, specs = "group")
 group              emmean   SE df lower.CL upper.CL
 Chemistry student    38.6 3.24 41     32.0     45.1
 Math student         40.5 1.60 41     37.2     43.7
 Physics student      39.0 2.22 41     34.5     43.5

Confidence level used: 0.95
```

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")
```