MET CS 555 Assignment 5 – 20 points

Fall 2, 2019

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. You will also lose 10 points per day for late submissions unless prior arrangements are made with your facilitator.

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

My answer:

Each group has 15 students, and the summarization of data relating to the test scores and age by the student group are given below.

IQ scores by student group:

Group	Population	Mean of scores	Standard deviation
Chemistry student	15	46.26667	3.731462
Math student	15	37.6	5.526559
Physics student	15	34.13333	4.657815

Age by student group:

Group	Population	Mean of age	Standard deviation
Chemistry student	15	40.06667	4.216747
Math student	15	20.73333	2.987275
Physics student	15	17.13333	1.846490

(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

My answer:

After performing one way ANOVA using the aov function in R, we get the result:

	Df	Sum sq	Mean sq	F-value	p-value
grp	2	1171.7	585.9	26.57	3.5e-08
Residuals	42	926.3	22.1		

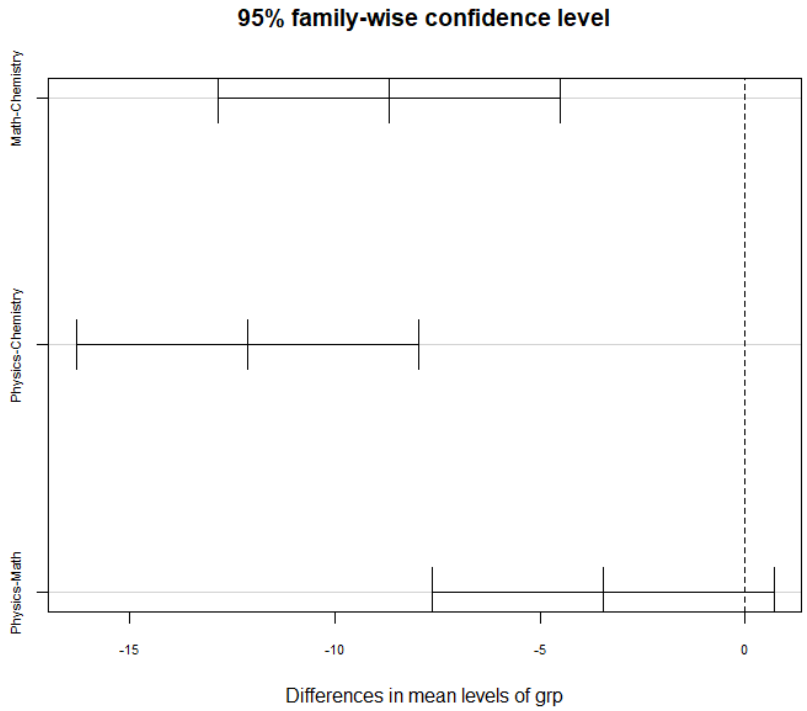
Looking at the F-test, the p-value is much less than α=0.05, and the F-value = 26.57 is greater than 3.22 (the F-value with degree of freedom at 2 and 42), thus, we reject the null hypothesis. We have significant level at the α=0.05 level that there is a difference in IQ test scores vary by student group.

Then, we need to perform the pairwise comparisons to adjust for multiple comparisons. Here we set “Chemistry student” as reference group, and we can get the result in table and plot below.

Tukey multiple comparisons of means (95% family-wise confidence level):

	mean difference	lwr	upr	p adj
Math - Chemistry	-8.666667	-12.83	-4.50	0.0000262
Physics - Chemistry	-12.133333	-16.29	-7.96	0
Physics - Math	-3.466667	-7.63	0.69	0.1194835

Graph of 95% family-wise confidence level:



As we seen from the plot of confidence interval of the pairwise comparisons, we found that there's not being a significant difference between the physics and math students, since the 0 is included in its confidence interval, that is, we are 95% sure that the true mean difference in IQ test score between physics and math students is between the lower bound, and its upper bound (from -7.63 to 0.69), and therefore, it might be no difference at all because it includes zero.
Furthermore, both the p-value and confidence interval of the other 2 pair of groups (physics - chemistry, math - chemistry) prove that they have significant level at the α=0.05 that there is a difference in IQ test scores vary by student group.

(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. What is the interpretation of the beta estimates from the regression model? – 4 points

My answer:
Create dummy variables and one way ANOVA using lm function in R. Here we set g0 as the group of chemistry student (as reference group), g1 for math student and g2 for physics student.

```
data$g0 <- ifelse(data$group == "Chemistry student", 1, 0)
data$g1 <- ifelse(data$group == "Math student", 1, 0)
data$g2 <- ifelse(data$group == "Physics student", 1, 0)

mche <- lm(iq~g1+g2, data = data)
summary(mche)
```

Result:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.267      1.213  38.157  < 2e-16 ***
g1            -8.667      1.715  -5.054 8.93e-06 ***
g2           -12.133      1.715  -7.076 1.13e-08 ***

F-statistic: 26.57 on 2 and 42 DF,  p-value: 3.496e-08
```

Thus, we can confirm that the result given above is the same as what we get from the one-way ANOVA using aov function in question 2.
In this case, we have three beta estimate value. The beta estimate value of intercept β_0 is the sample mean in the reference group, which is the average score of chemistry student, is 46.267. The beta estimate value of β_1 (coefficient of g1) is the mean difference between group of Math student and Chemistry student, which is -8.667. At

last, the beta estimate value of β_2 (coefficient of g2) is the mean difference between group of Physics student and Chemistry student, which is -12.133.

(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 questions it answers as opposed to the one above). Did you obtain different results? Summarize briefly (no need to go through the 5 –step procedure here). Present the least square means and interpret these. – 6 points

My answer:

By re-do the one-way ANOVA adjusting for age, we get the different result from the analysis in question 2, the result is:

```
            Sum Sq Df F value   Pr(>F)
(Intercept) 152.74  1  7.8294 0.007797 **
grp          21.89  2  0.5610 0.574969
age         126.42  1  6.4804 0.014763 *
Residuals   799.84 41
```

That is, after adjusting for age, the differences seen in the previous one-way ANOVA setting were attenuated and the F-test for the effect of IQ test score in group was no longer significant (F = 0.561 on 2 and 41 degrees of freedom, p = 0.575). The p-value becomes huge and much greater than 0.05.

Get the least square means:

```
library(lsmeans)
options(contrasts=c("contr.treatment","contr.poly"))
m_age <- lm(iq~grp+age, data = data)
lsmeans(m_age, pairwise ~ grp, adjust = "none")
```

Get the result:

```
$lsmeans
 grp       lsmean   SE df lower.CL upper.CL
 Chemistry   38.6 3.24 41     32.0     45.1
 Math        40.5 1.60 41     37.2     43.7
 Physics     39.0 2.22 41     34.5     43.5

Confidence level used: 0.95

$contrasts
 contrast           estimate   SE df t.ratio p.value
 Chemistry - Math     -1.920 4.46 41 -0.430  0.6691
 Chemistry - Physics  -0.425 5.19 41 -0.082  0.9352
 Math - Physics        1.495 1.79 41  0.836  0.4081
```

The least square means (adjust for age) were 38.6, 40.5 and 39.0 for the chemistry, math and physics students, respectively. The difference between each least square

means seems small, and if we step further to look at the summary of data in age (already listed in the first question),

Group	Population	Mean of age	Standard deviation
Chemistry student	15	40.06667	4.216747
Math student	15	20.73333	2.987275
Physics student	15	17.13333	1.846490

We found that mean of age from three groups differ tremendously from group to group, as such, the difference that we saw in the one-way ANOVA model were due to the age differences across the student groups as opposed to true differences in IQ test score attributes only to their major. Based on the illustration above, when we take the age into consideration and adjustment, the difference of IQ test scores between 3 groups was no more significant.

R script:

```
setwd("C:/Users/Lin/Desktop/2019 Fall/555 DAV_ R/Homework/hw5")


data <- read.csv("IQtest.csv")
library(plyr)
ddply(data, "group", summarise,
        N = length(iq),
        mean = mean(iq),
        sd = sd(iq))


ddply(data, "group", summarise,
    N = length(age),
    mean = mean(age),
    sd = sd(age))


boxplot(iq~group, data = data, main = "xxx", xlab = "Group",
     ylab = "IQ")


data$grp <- revalue(data$group, c("Chemistry student" = "Chemistry",
              "Math student" = "Math",
              "Physics student" = "Physics"))


# one-way ANOVA using aov function
m <- aov(iq~grp, data = data)
summary(m)
```

```
# Pairwise comparison using Tukey's procedure
TukeyHSD(m)
plot(TukeyHSD(m), cex.axis = .7)
plot(TukeyHSD(m), cex.axis = .7, las = 2)

# create dummy variables
data$g0 <- ifelse(data$group == "Chemistry student", 1, 0)
data$g1 <- ifelse(data$group == "Math student", 1, 0)
data$g2 <- ifelse(data$group == "Physics student", 1, 0)

# One way ANOVA using lm function
mche <- lm(iq~g1+g2, data = data)
summary(mche)

# adjusting for age
library(car)
Anova(lm(iq~grp+age, data = data), type = 3)
Anova(lm(iq~grp, data = data), type = 3)

# Generate Least Squares means and comparisons
library(lsmeans)
options(contrasts=c("contr.treatment","contr.poly"))
m_age <- lm(iq~grp+age, data = data)
lsmeans(m_age, pairwise ~ grp, adjust = "none")
lsmeans(m_age, pairwise ~ grp, adjust = "tukey")
```