Dhakal_Module5

Kokil Dhakal

2023-12-04

Loading Dataset

```
data.module5 <- read_excel("/Users/kokildhakal/Desktop/STUDY/BU/9.CS555/Module5/module5.data.xlsx")
head(data.module5)
```

```
## # A tibble: 6 x 3
##   group              iq   age
##   <chr>           <dbl> <dbl>
## 1 Physics student    34    15
## 2 Physics student    33    17
## 3 Physics student    32    15
## 4 Physics student    25    14
## 5 Physics student    36    19
## 6 Physics student    30    18
```

```
#checking group variable as factor or not
#is.factor(data.module5$group)
#since group is not factor, converting it to factor
#as.factor(data.module5$group)
#cheking datatypes of the dataset.
str(data.module5)
```

```
## tibble [45 x 3] (S3: tbl_df/tbl/data.frame)
##  $ group: chr [1:45] "Physics student" "Physics student" "Physics student" "Physics student" ...
##  $ iq   : num [1:45] 34 33 32 25 36 30 31 34 29 34 ...
##  $ age  : num [1:45] 15 17 15 14 19 18 16 17 16 17 ...
```

```
attach(data.module5)
```

Question1

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.

Answer:

```r
#calculating number of students in each group
no.of.student <- table(group)
kable(no.of.student,col.names = c("Type of student","Number of student"),align = "c")
```

Type of student	Number of student
Chemistry student	15
Math student	15
Physics student	15

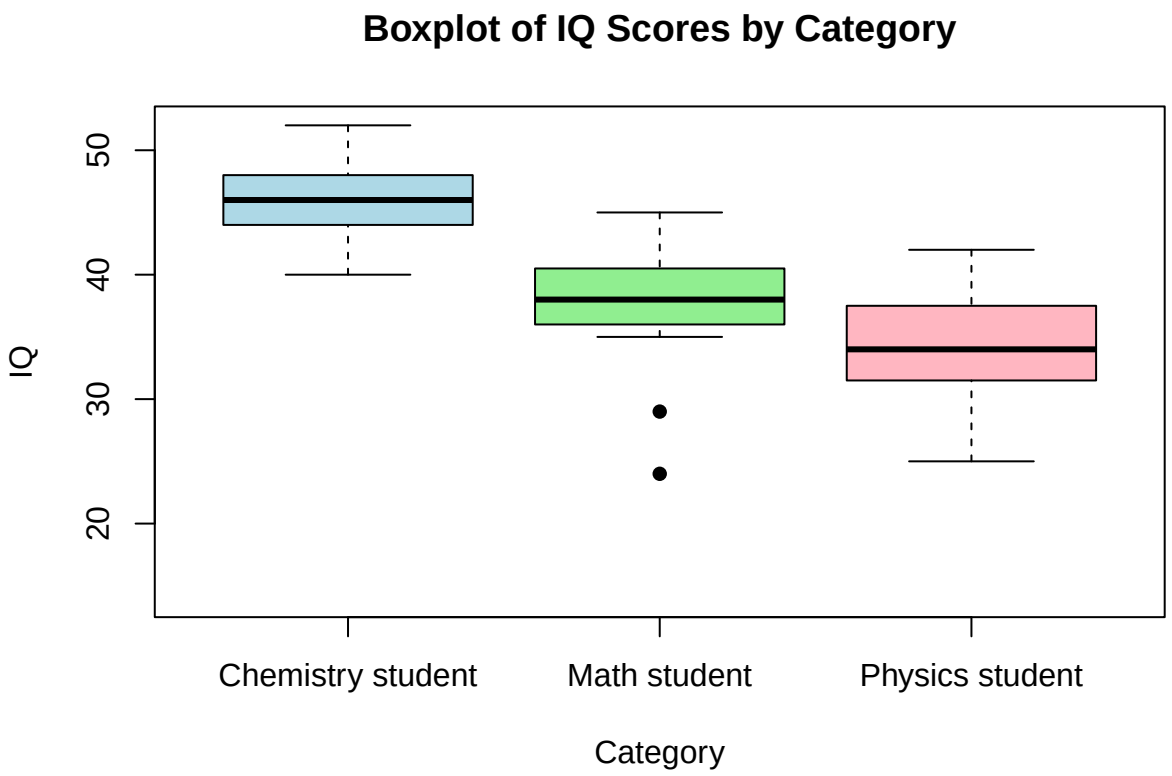
```r
#making dataframe with sd for each group
iq.sd <- data.frame(as.matrix(aggregate(iq, by=list(group), sd)))
colnames(iq.sd) <- c("group","sd")
age.sd <-data.frame(as.matrix(aggregate(age, by=list(group),sd)))
colnames(age.sd) <- c("group","sd")

#making summary for each goup inlcluding sd

iq.summary <-aggregate(iq, by=list(group), summary)
age_summary <- aggregate(age,by=list(group),summary)
iq.df <- data.frame(as.matrix(iq.summary))
iq.df <- cbind(iq.df,iq.sd[,2])
colnames(iq.df) <- c("group","min","Q1","median","mean","Q3","max","sd")
kable(iq.df,align = "c",label = "test scores")
```

group	min	Q1	median	mean	Q3	max	sd
Chemistry student	40.00000	44.00000	46.00000	46.26667	48.00000	52.00000	3.731462
Math student	24.00000	36.00000	38.00000	37.60000	40.50000	45.00000	5.526559
Physics student	25.00000	31.50000	34.00000	34.13333	37.50000	42.00000	4.657815

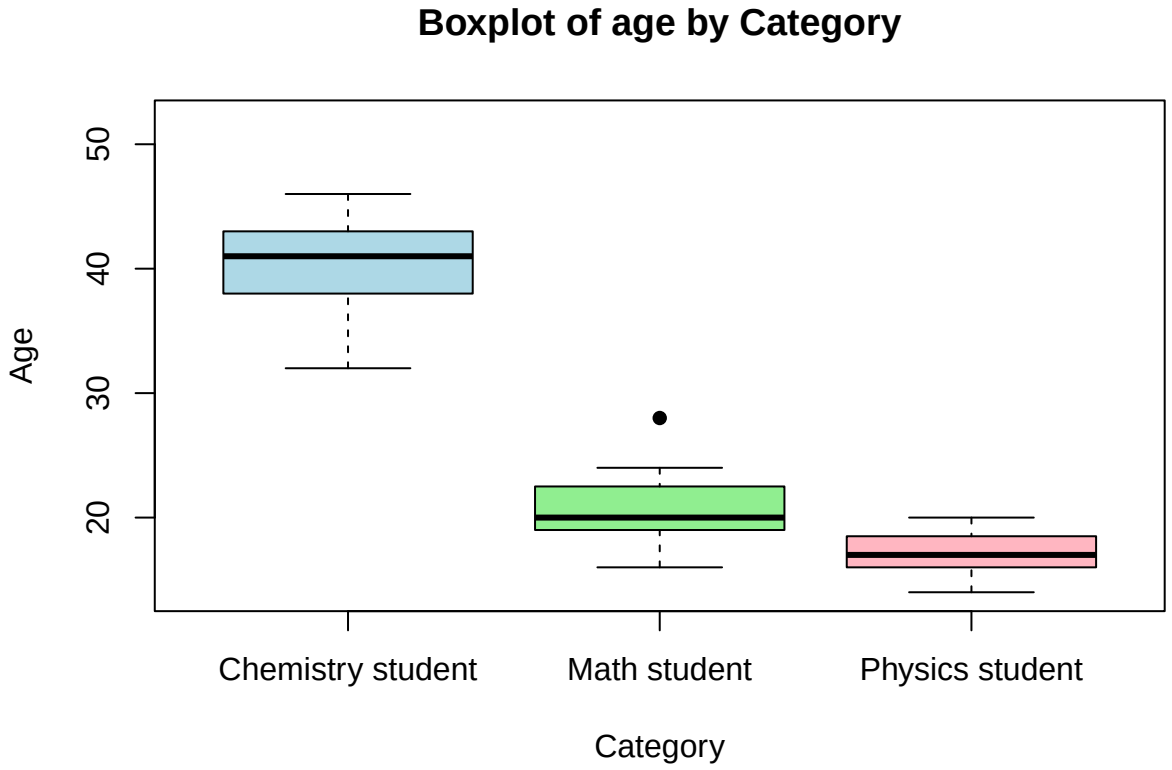
```r
#calculating summary by age
age.df <- data.frame(as.matrix(age_summary))
age.df <- cbind(age.df,age.sd[,2])
colnames(age.df) <- c("group","min","Q1","median","mean","Q3","max","sd")
kable(age.df,align = "c",label = "Age")
```

group	min	Q1	median	mean	Q3	max	sd
Chemistry student	32.00000	38.00000	41.00000	40.06667	43.00000	46.00000	4.216747
Math student	16.00000	19.00000	20.00000	20.73333	22.50000	28.00000	2.987275
Physics student	14.00000	16.00000	17.00000	17.13333	18.50000	20.00000	1.846490

```r
y_limits <- range(c(age,iq))
boxplot(iq ~ group, data = data.module5,
        main = "Boxplot of IQ Scores by Category",
        xlab = "Category",
        ylab = "IQ",
        col = c("lightblue", "lightgreen", "lightpink"),
        ylim=y_limits,
        pch=16)
```




```r
boxplot(age ~ group, data = data.module5,
        main = "Boxplot of age by Category",
        xlab = "Category",
        ylab = "Age",
        col = c("lightblue", "lightgreen", "lightpink"),
        ylim=y_limits,
        pch=16)
```



First step is to load the dataset into the R environment and then checking group variable whether it is a factor or not.If it is not it is converted using as.factor() method.After that table() method is used to check number of student present in each group.it is found that each group has 15 students.After that calculating sd for each variables age and test for each group (which is later combined with summary table).After that it summarized the results using aggregate function for both age and test scores by the student group.I made Summarized table for both age and test scores as above which includes min,Q1,median,mean,Q3,max and sd values for age and test scores for all three student types.Also, I made box plot for both age and test scores.It seems from box plot that there are some outliers in both age and test scores variables for Math student group.From summary and plot we can see that mean age of chemistry student is higher followed by math student.Similarly, Mean test scores of chemistry student is higher followed by math student.Similarly, sd of test scores vairable is higher for math student followed by physics and then chemistry.In case of age variable, sd of age is higher for chemistry followed by math and then physics.

Question2

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 =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.

Answer: Performing one way ANOVA

```
one.way.analysis<- aov(iq~group,data=data.module5)
summary(one.way.analysis)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
```

```
## group        2 1171.7   585.9   26.57 3.5e-08 ***
## Residuals   42  926.3    22.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now using 5-steps procedure for hypothesis testing

1 Setting up the hypotheses and selecting alpha level

- H0: $\mu(ChemistryStudent)$= $\mu(MathStudent)$=$\mu(PhysicsStudent)$ (All underlying population means for test score are equal)
- H1: $\mu i \neq \mu j$ for some i and j.(Not all of the underlying population means for test scores are equal)
- $\alpha = 0.05$

2 Selecting the appropriate test statstic

- F= Mean Squares(between)/Mean squares (within)

3 Decision rule:

- Determine the appropriate value from the F-distribution with k−1=3−1=2, n−k=45−3=42 degrees of freedom and associated with a right hand tail probability of alpha=0.05
- Using the software, F(k−1, n−k,)

```
qf(0.05,2,42,lower.tail = FALSE)
```

```
## [1] 3.219942
```

- Decision Rule: Reject H0 if F 3.2199
- Otherwise, do not reject H0

4 Compute the test statistic

- F= Mean Squares(between)/Mean squares within =585.9/22.1 = 26.57

5 Conclusion

- Since F-statistic is greater than critical value at level 0.05,we reject Null hypotheses.This means Not all of the underlying population means for test scores are equal.This also means that there is difference in mean test score between at least one pair group out of three(math-chemistry,math-physics,chemistry-physics).

Now performing pairwise comparison using tukey's procedure

```
one.way.tukey <- TukeyHSD(one.way.analysis,conf.level = 0.95)
tukey.table <- one.way.tukey$group
kable((tukey.table),digits = 4,align = "c")
```

	diff	lwr	upr	p adj
Math student-Chemistry student	-8.6667	-12.8328	-4.5006	0.0000
Physics student-Chemistry student	-12.1333	-16.2994	-7.9672	0.0000
Physics student-Math student	-3.4667	-7.6328	0.6994	0.1195

First I performed global F-test.Based on critical value and F-statstic, reject null hypothesis which means Not all of the underlying population means for test scores are equal.However we do not know mean test scores between which group pair.It could be all three pair, it could be two pair or just one pair. For this we performed Tukey's procedure which gives p- values for corresponding pair(pairwise).From above table, where adjusted p-values using Tukey's method shows there is very low p-values for math student vs Chemistry student which means at level of 0.05, there is difference in mean of test scores between Math student and Chemistry student. Also, there is very low p-value for group physics student and chemistry student which means at level of 0.05, there is difference in mean of test scores between physics student and chemistry student.However, since p-value is higher than alpha value(0.05) for the student group between physics student and math student,we can not reject null hypothesis that says mean test score between physics and math student is equal.Also, if we see 95% confidence interval,0 is within the interval levels(-7.632 and 0.6994) which means difference in mean test scores between math student and physics student is 0. This calculation shows test scores vary by student group.

Question 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 the results are the same (specifically point out test statistics, p-values, etc. that show the results are equivalent). What is the interpretation of the beta estimates from the regression model?

Answer:

Since chemistry students is our reference group.while making dummy variable, particular group will be assigned to 1 and all other group will be assigned as 0.

```
data.module5$g0 <- ifelse(data.module5$group =="Chemistry student",1,0)
data.module5$g1 <- ifelse(data.module5$group =="Physics student",1,0)

data.module5$g2 <- ifelse(data.module5$group =="Math student",1,0)
```

Now setting up linear regression and chemistry group as reference group

```
linear_regression_model <- lm(iq~g1+g2,data = data.module5)
summary(linear_regression_model)
```

```
##
## Call:
## lm(formula = iq ~ g1 + g2, data = data.module5)
##
## 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 ***
## g1           -12.133      1.715  -7.076 1.13e-08 ***
## g2            -8.667      1.715  -5.054 8.93e-06 ***
## ---
## 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
```

Here, after making dummy variables I set up multiple linear regression(chemistry student group is reference group) model.To compare the ANOVA and regression model, iq variable will be the repose variable for regression model as this same variable is used in our previous one-way ANOVA. Values of summary stat of both linear regression and ANOVA looks same. Global F statistic from both calculation is 26.57 with 2 and 42 df and similar p-values.Here intercept of the regression model is equal to the mean of Chemistry student test scores(reference group). and Estimate of g1 and g2 is equal to the difference of means of physics student test scores and math student test with mean chemistry student test scores respectively.

from above table our regression model will be

- $y = 46.27 - 12.133 * g1 - 8.667 * g2$

Here, Slope of regression line is basically difference of mean of a particular group with mean score of reference group that is beta(g1) is difference of mean test scores between physics and chemistry student. and beta(g2) is difference of mean test scores between Math and chemistry student.

Question4

Re-do the one-way ANOVA adjusting for age (ANCOVA). 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). Lastly, present the least square means and interpret these.

Answer:

Now running ANOVA with adjusting for age

```
Anova(lm(iq ~ group + age),type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: iq
##             Sum Sq Df F value   Pr(>F)
## (Intercept) 152.74  1  7.8294 0.007797 **
## group        21.89  2  0.5610 0.574969
## age         126.42  1  6.4804 0.014763 *
## Residuals   799.84 41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now calculating least mean squares

```
module.5.model <- lm(iq~ group+age,data = data.module5)
emm_options(contrasts=c("contr.treatment","contr.apply"))
emmeans(module.5.model,specs = "group",contr = "pairwise")
```

```
## $emmeans
##  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
##
## $contrasts
##  contrast                              estimate   SE df t.ratio p.value
##  Chemistry student - Math student        -1.920 4.46 41  -0.430  0.9031
##  Chemistry student - Physics student     -0.425 5.19 41  -0.082  0.9963
##  Math student - Physics student           1.495 1.79 41   0.836  0.6832
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

From above ANCOVA with adjusted age, it seems from pairwise comparison that there is no difference in mean iq score between variables as p value between any pair is greater than alpha value(i.e 0.05)Which means we can not reject null hypothesis,which states, all underlying population means for test scores are equal. In the step 2 where we did not control age covariate, there is difference mean test scores between physics student and chemistry student and, math student and chemistry student where p-value is less than alpha value however there was no significance difference of means of test scores between physics and math students.Also, global F-statstic without adjusting age gives 26.57 with p value less than 0.05 (alpha value) this states that not all of the underlying population means for test scores are equal. However after adjusting age if we calculate F-stat,it gives 0.56 with p-value greater than 0.05 (alpha level) which can be interpreted as all underlying population mean of test scores in different student group are equal.However, p-value of covariate age itself show less than alpha value.This can be interpreted as the differences that we saw in the step 2 using one-way anova model were due to age differences across the test scores groups as opposed to true difference in mean test scores attributable only to type of student groups.So there is different result obtained after adjusting age variable.Lastly, when we check the least square means for each group after adjusting age, it is found that chemistry student has 38.6, math student has 40.5 and physics student has 39.0.The pairwise comparison of these means show that there is not much difference in mean of test scores between different student group.
