CS555, Data Analysis and Visualization Homework 5

Jefferson Parker, [japarker@bu.edu](mailto:japarker@bu.edu)

See the accompanying R code at the end of this document.

1. For the data set provided, there are 15 students per group (data not shown). Overall, the chemistry students have generally had higher age (Table 1) and IQ (Table 2) than math and physics students as shown in both the summary tables and box plots (Figure 1, Figure 2) below.



Figure 1: Box plot of Age by student group (left).

Figure 2: Box plot of IQ by student group (right).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | IQ | | | | | |
| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| Chemistry | 40 | 44 | 46 | 46.26667 | 48 | 52 |
| Math | 24 | 36 | 38 | 37.6 | 40.5 | 45 |
| Physics | 25 | 31.5 | 34 | 34.13333 | 37.5 | 42 |

Table 1: Summary statistics of IQ by student group.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Age | | | | | |
| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| Chemistry | 32 | 38 | 41 | 40.06667 | 43 | 46 |
| Math | 16 | 19 | 20 | 20.73333 | 22.5 | 28 |
| Physics | 14 | 16 | 17 | 17.13333 | 18.5 | 20 |

Table 2: Summary statistics of Age by student group.

As seen in the box plots, the variability in the IQ data is wider than the age data.

1. To test if the IQ scores vary by group we use one-way ANOVA.
   1. Our hypothesis is that the group wise mean IQ scores do not differ:

H0: chem = math = phys

H1: 1 ≠ 2 for some group 1 and 2

 = 0.05

* 1. Our test statistic is with dfMSB = k-1 = 2, dfMSW = n-k = 42 and  = 0.05 which has a value of **3.22**.
  2. We fail to reject H0 if the F statistic is greater than our critical F-value of 3.22, or if the p-value is less than our critical value  = 0.05. Otherwise we reject H0.
  3. Our test statistic is F = 26.57, with a p-value of 3.5-8.
  4. By either of the test statistics we reject the null hypothesis and conclude that there is a difference in IQ test result between at least two of the student groups.

Given that the global F-test determined that there is an actual difference in IQ scores among at least two of the student groups, we can perform pairwise comparisons to determine just which groups are different.

**NOTE** – the assignment said to do a pairwise comparison using Tukey’s procedure for multiple comparison adjustment, however the pairwise.t.test function does not accept ‘Tukey’ as an adjustment parameter. Instead I have performed the pairwise comparison using the pairwise t-test and Bonferroni’s adjustment AND Tukey’s HSD. Both returned similar results.

The pairwise test of differences in sample mean IQ’s between student groups determined that there were significant differences between Chemistry student’s and the other groups (Table 3):

|  |  |  |  |
| --- | --- | --- | --- |
| Group 1 | Group 2 | Difference | p-valueadj |
| Math | Chemistry | -8.7 | 2.7-5 |
| Phyiscs | Chemistry | -12.1 | 3.4-8 |

Table 3: Pairwise t-test results of IQ between student groups.

1. Assigning dummy variables to the group levels and re-running the one-way ANOVA as a linear regression (with Chemistry students as the reference level) generates the same differences but different p-values (Table 4):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |  |
| (Intercept) | 42.267 | 1.213 | 38.157 | < 2e-16 | \*\*\* |
| gMath | -8.667 | 1.715 | -5.054 | 8.93E-06 | \*\*\* |
| gPhys | -12.133 | 1.715 | -7.076 | 1.13E-08 | \*\*\* |

Table 4: Linear regression results of One-Way ANOVA of pairwise differences.

In this case, the results of the beta estimates indicate that *on average* Math students score 8.67 points lower than Chemistry students on the IQ test, while Physics students *on average* score 12.13 points lower than Chemistry students.

1. Repeating the one-way ANOVA adjusting for age we find that age is a significant factor, while group membership actually is not significant (Table 5). Performing a pairwise comparison of IQ test means between student groups, after controlling for age, shows very different results from the results in Table 4 (Table 6).

When the IQ scores among student groups are compared for differences, when controlling for age, we find that there is no evidence of a difference by p-value. In fact, the difference in scores is actually explained by the difference in the average age of students in the different groups.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| (Intercept) | 152.74 | 1 | 7.8294 | 0.007797 |
| group | 21.89 | 2 | 0.561 | 0.574969 |
| age | 126.42 | 1 | 6.4804 | 0.014763 |
| Residuals | 799.84 | 41 |  |  |

Table 5: ANCOVA of IQ score against student group and age.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| contrast | estimate | SE | df | t.ratio | p.value |
| Chemistry student - Math student | -1.9201982 | 4.46057 | 41 | -0.43 | 0.9031 |
| Chemistry student - Physics student | -0.4248788 | 5.190133 | 41 | -0.082 | 0.9963 |
| Math student - Physics student | 1.4953194 | 1.789081 | 41 | 0.836 | 0.6832 |

Table 6: Pairwise comparison of IQ score among student groups, controlling for age.

As first noted in Figure 1, boxplots of age by group, the Chemistry students were generally older than physics and math students. This age difference could translate into an advantage on IQ test performance among the Chemistry student group.

**R Code**

# CS555 Data Analysis and Visualization

# Homework5.R

# Jefferson Parker, japarker@bu.edu

# 20180803

# Load libraries.

library(car);

library(lsmeans);

# Save the student data to a file and load to R.

inputDir <- "C:/Users/jparker/Code/Input";

setwd(inputDir);

studentData <- read.table(file = "studentIq.txt", header = TRUE, sep = "\t");

# 1. How many students are in each group.

# Summarize the data relating to both test score and age by group.

aggregate(studentData, by = list(studentData$group), summary);

boxplot(age ~ group, data = studentData, main = "Age per Student Group");

boxplot(iq ~ group, data = studentData, main = "IQ per Student Group");

# 2. Do test scores vary by group?

# Critical F value

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

testModel <- aov(iq ~ group, data = studentData);

summary(testModel);

# If the overall model is significant, perform pairwise testing.

# Note of confusion. The homework says use Tukey's adjustment method but that is

# not an option for pairwise.t.test. I am using both pairwise.t.test and TukeyHSD

# to check both methods.

pairwise.t.test(studentData$iq, studentData$group, p.adjust.method = 'bonferroni');

TukeyHSD(testModel);

# 3. Create the appropriate number of dummy variables for student group

# and re-run the one way ANOVA using the lm function.

# Set 'Chemistry student' as the reference group.

studentData$gC <- ifelse(studentData$group == 'Chemistry student', 1, 0);

studentData$gM <- ifelse(studentData$group == 'Math student', 1, 0);

studentData$gP <- ifelse(studentData$group == 'Physics student', 1, 0);

lineModel <- lm(iq ~ gM + gP, data = studentData);

summary(lineModel);

# 4. Re-do the one way ANOVA adjusting for age (ANCOVA).

Anova(lm(iq ~ group + age, data = studentData), type = 3);

# set our categorical variable options.

options(contrasts = c("contr.treatment", "contr.poly"));

# Measure pairwise differences between groups, accounting for differences in age.

lsmeans(lm(iq ~ group + age, data = studentData), pairwise ~ group, adjust = 'Tukey');

# Just checking

# install.packages("emmeans");

library(emmeans);

emmeans(lm(iq ~ group + age, data = studentData), pairwise ~ group, adjust = 'Tukey');