MET CS555: Homework 5

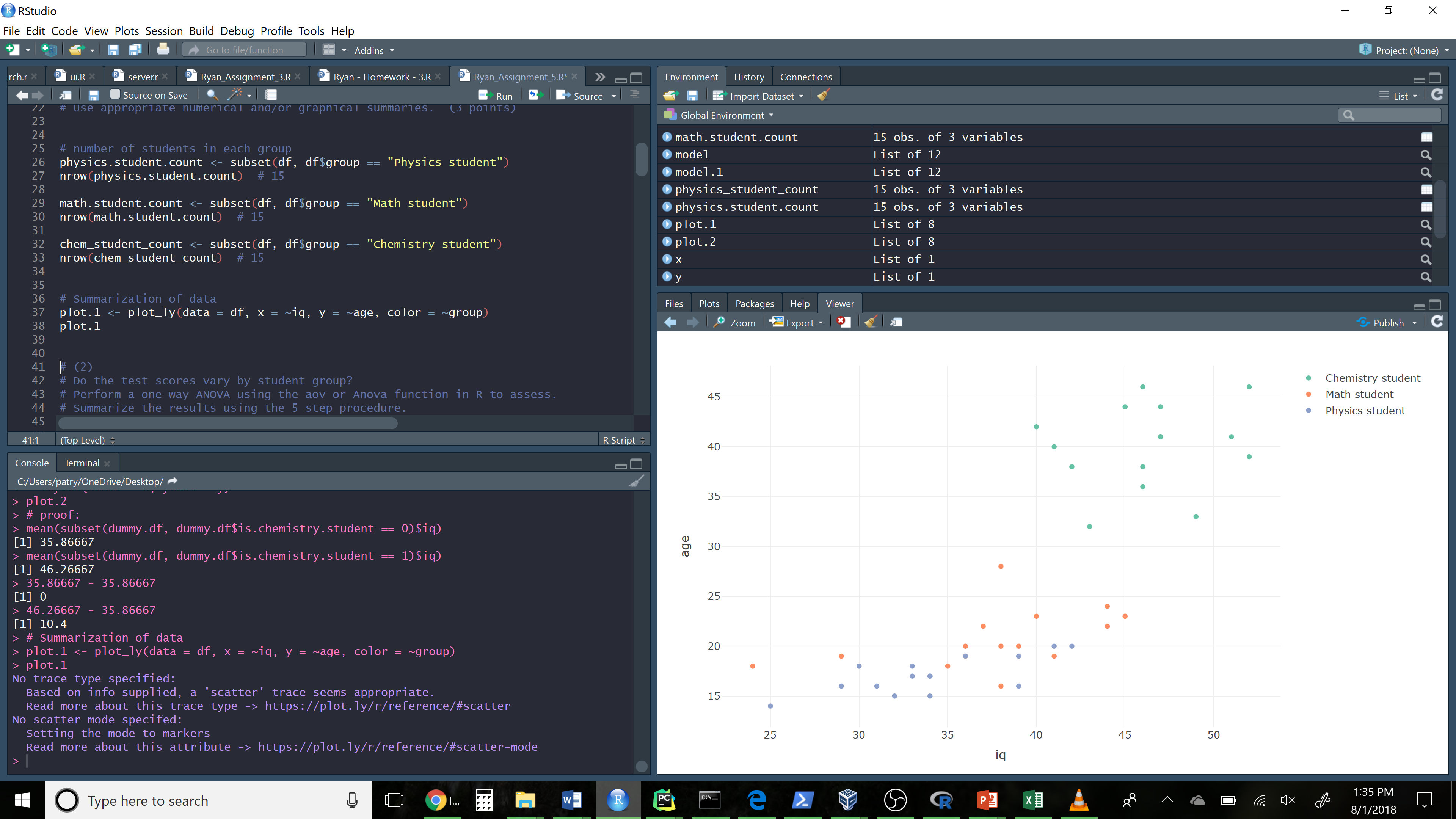
Patrick Ryan



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)

|  |
| --- |
| # (1) Save the data to a file (excel or CSV file) and read it into R memory for analysis. (Q1 - 2 points)  setwd("C:/Users/patry/OneDrive/Desktop")  df = read.csv("student\_iq\_test.csv", header=T, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  df  # 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)  # number of students in each group  physics.student.count <- subset(df, df$group == "Physics student")  nrow(physics.student.count) # 15  math.student.count <- subset(df, df$group == "Math student")  nrow(math.student.count) # 15  chem\_student\_count <- subset(df, df$group == "Chemistry student")  nrow(chem\_student\_count) # 15  # Summarization of data  plot.1 <- plot\_ly(data = df, x = ~iq, y = ~age, color = ~group)  plot.1 |

[Figure 1: 2-Dimensional Scatter Plot of Age vs IQ]



1. Do the test scores vary by student group? Perform a one way ANOVA using the aov or Anova function in R to assess. 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)

|  |
| --- |
| # Analysis of variance model  model <- aov(iq ~ group, data=df)  summary(model)  anova.table <- anova(model)  anova.table  # Analysis of Variance Table  #  # Response: iq  #  # Df Sum Sq Mean Sq F value Pr(>F)  # group 2 1171.73 585.87 26.565 3.496e-08 \*\*\*  # Residuals 42 926.27 22.05  # (assume alpha of 0.05)  # (1)  # H0: All underlying population means are equal  # H1: Not all of the underlying population means are equal  # (2) Select the appropriate test statistic  # F = MSB / MSW  # (3)  # State the decision rule:  # Decision rule: Reject H0 if f >= 3.219942  # Otherwise reject H0:  k <- 3  n <- 45  qf(0.95, df1= k - 1, df2= (n - k)) # 3.219942  # (4) Compute the test statistic  MSB <- anova.table$`Mean Sq`[1]  MSW <- anova.table$`Mean Sq`[2]  f = MSB / MSW  f #26.56514  #(5) Conclusion  # Reject H0 since 26.56 is >= 3.2199  # Significance in findings confirmed  # Analysis via TukeyHSD  model <- aov(iq ~ group, data=df)  TukeyHSD(model)  # SUMMARIZED RESULT  # $`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  model.1 <- model  model.1 |

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

|  |
| --- |
| # create appropriate dummy variables  dummy.df <- df  dummy.df$is.chemistry.student <- ifelse(df$group == "Chemistry student", 1, 0)  dummy.df$is.math.student <- ifelse(df$group == "Math student", 1, 0)  dummy.df$is.physics.student <- ifelse(df$group == "Physics student", 1, 0)  dummy.df  model <- lm(dummy.df$iq ~ dummy.df$is.physics.student + dummy.df$is.math.student, data = dummy.df)  summary(model)  model.2 <- model  # Confirm if the results are the same:  anova(model.1)  anova(model.2)  # Proof:  # model.1's residual Sum Sq (926.27) == model.2's Residual Sum Sq (926.27) (same applies for Mean Sq)  # model.1's group Sum Sq (117.73) == model.2's variable Sum Sq (dummy.df$is.physics.student + dummy.df$is.math.student)  # model.1's F value (26.565) is the average of the two F values in model.2, i.e. (27.587 + 25.543)/2  # model.1's Mean Sq is the average of model.2's variable Mean Sq coefficients (608.40 + 563.33)/2  summary(model.2)  # Visualize B1  x <- list(title = "Is Chemistry Student")  y <- list(title = "IQ Test Score")  plot.2 <- dummy.df %>%  plot\_ly(x = dummy.df$is.chemistry.student) %>%  add\_markers(y = dummy.df$iq, name="Observation") %>%  add\_lines( y = fitted(model), name="Beta.1 Projection") %>%  layout(xaxis = x, yaxis = y)  plot.2  summary(model)  # Interpretation:  # The y intercept looks to be the mean of the students who are not chemistry students  # and it appears to slope to meet the mean of students who are chemistry students  # proof:  mean(subset(dummy.df, dummy.df$is.chemistry.student == 0)$iq)  mean(subset(dummy.df, dummy.df$is.chemistry.student == 1)$iq)  # This logically implies that a linear model assumes an increase of 10.4 iq points between chemistry and non chemistry students  # 46.26667 - 35.86667 = 10.4 |

Yes, the models are the same:

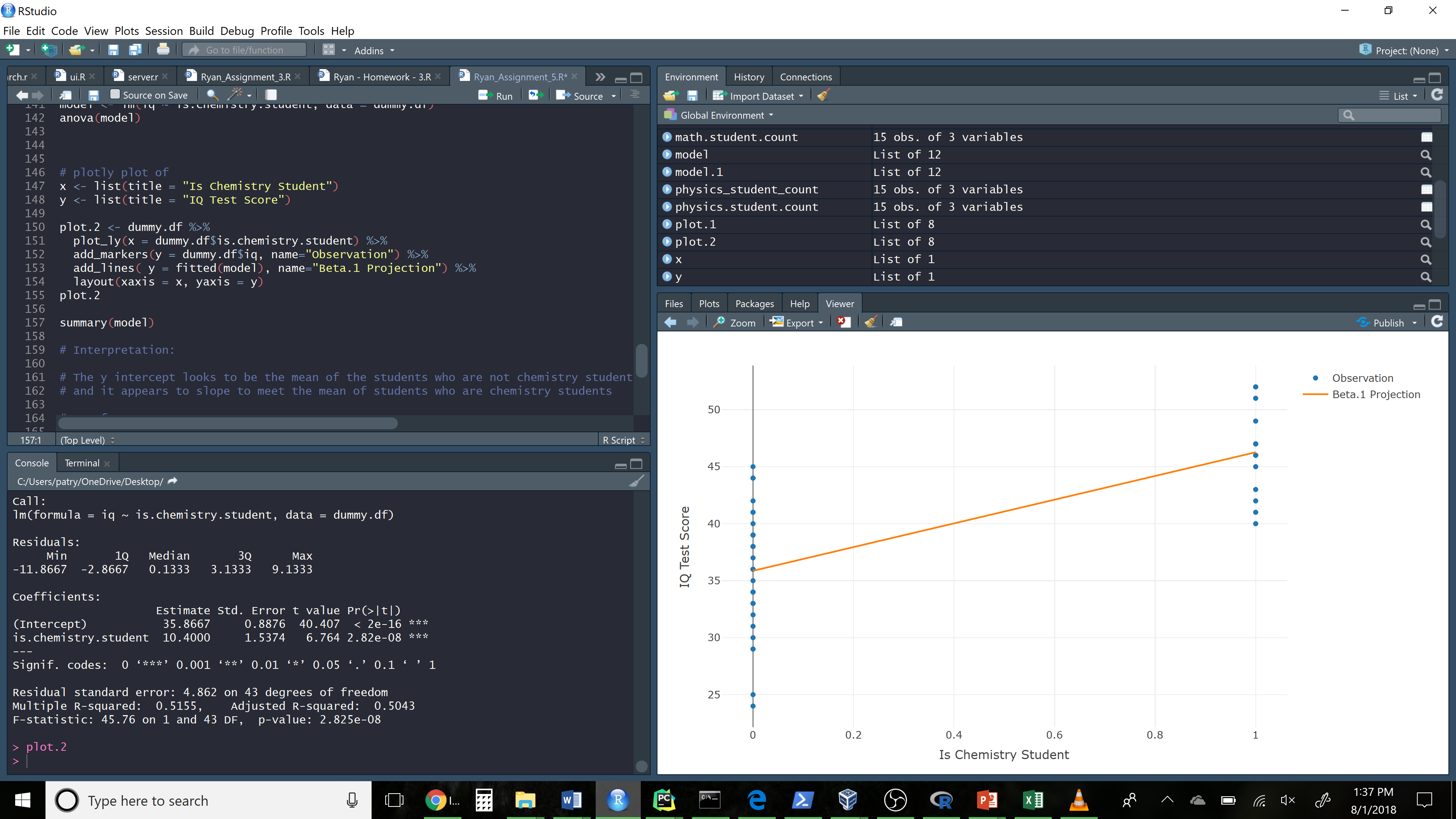
Proof:

* model.1's residual Sum Sq (926.27) == model.2's Residual Sum Sq (926.27) (same applies for Mean Sq)
* model.1's group Sum Sq (117.73) == model.2's variable Sum Sq (dummy.df$is.physics.student + dummy.df$is.math.student)
* model.1's F value (26.565) is the average of the two F values in model.2, i.e. (27.587 + 25.543)/2
* model.1's Mean Sq is the average of model.2's variable Mean Sq coefficients (608.40 + 563.33)/2

Interpretation ([see Figure 2]):

* The y intercept looks to be the mean of the students who are not chemistry student sand it appears to slope to meet the mean of students who are chemistry students
* This logically implies that a linear model assumes an increase of 10.4 iq points between chemistry and non chemistry students

[Figure 2: Fitting of Linear Model with Dummy Variables]



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 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)

|  |
| --- |
| # plot means using plotly boxplot  plot.3 <- plot\_ly(df, y = ~iq, color = ~group, type = "box")  plot.3  model <- aov(iq ~ group + age, data=df)  anova(model)  Anova(lm(df$iq~df$group+df$age), type=3)  # How is this analysis different?  # This analysis accounted for adjustment of the age variable.  # This time we are evaluating whether the means of a dependent categorical variable (IQ Test Score) are equal across levels of a  # categorical independent variable (student type), while statistically controlling for the effects of a continuous variable (age).  options(contrasts=c("contr.treatment", "contr.poly"))  # Calculate the lsmeans  lsmeans(lm(df$iq~df$group+df$age), pairwise~df$group)  #Interpretation: the ls means are the group means after having controlled for age. |

LS means output:

$`lsmeans`

df$group lsmean SE df lower.CL upper.CL

Chemistry student 32.96513 1.229289 41 30.48253 35.44773

Math student 34.06032 1.140782 41 31.75646 36.36418

Physics student 32.96513 1.229289 41 30.48253 35.44773

Confidence level used: 0.95

$contrasts

contrast estimate SE df t.ratio p.value

Chemistry student - Math student -1.095193 0.4302201 41 -2.546 0.0384

Chemistry student - Physics student 0.000000 0.0000000 41 NaN NaN

Math student - Physics student 1.095193 0.4302201 41 2.546 0.0384

P value adjustment: tukey method for comparing a family of 3 estimates

How is this analysis different?:

* This analysis accounted for adjustment of the age variable.
* This time we are evaluating whether the means of a dependent categorical variable (IQ Test Score) are equal across levels of a categorical independent variable (student type), while statistically controlling for the effects of a continuous variable (age).
* This is in contrast to the first test where we sought to perform a one-way ANOVA to assess the difference in means between each sampling group (using the group as a categorical variable, and iq as a continuous variable)

Summarize:

* Yes, we observed significantly different results in the ANCOVA test. After adjusting for age, we can see that there is not a large statistical significance related to groups (student major). (shifting from 9.29e-09 to 0.574969 over the two test outputs)

LS Means Interpretation:

* interpretation: the ls means are the group means as calculated after having controlled for a covariate (which in this case would have been the variable of student’s age).

[Figure 3: Boxplot of Means]

