

ST 411/511 Homework 5

Pavan Sai Nallagoni

Summer 2022

Instructions

This assignment is due by 11:59 PM, August 9th on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile using the provide .Rmd (R Markdown) template.**

Note: Create a PDF by either compiling a PDF directly (via LaTeX) or from a Word Doc. Do not submit a PDF of the HTML output as they're cumbersome and difficult to read.

Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope (Failure to do so will result in point deductions). You should also write complete, grammatically correct sentences for your solutions.

Once you've completed the assignment, and before you submit it to Gradescope, you should read the document to ensure that (1) The computed values show up in the document, (2) the document "looks nice" (i.e. doesn't have extraneous code/outputs and includes *just* the essentials), and (3) ensure the document is "easy" to read.

Goals:

1. Practice computing the different "pieces" we need to compute to conduct an ANOVA test.
2. Demonstrate the connections between ANOVA as a test of population means and as a method for model comparison.
3. Get hands on experience conducting ANOVA and Kruskal-Wallis tests.
4. Practice using linear combinations of means to answer scientific questions of interest.
5. Learn how to perform different methods of multiple comparison procedures with R.

Question 1 (7 points)

The table below shows a partially completed ANOVA table. (Note: if you are looking at this in RStudio it may be helpful to knit the file to properly view the table.)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic	p-value
Between Groups
Within Groups	35088	24	.		
Total	70907	31			

(a) (1 point) How many groups were there?

Degrees of Freedom for Within groups: $n - I = 24 \rightarrow \text{eq (1)}$

Degrees of Freedom for Total groups: $n - 1 = 31 \rightarrow \text{eq (2)}$

where n is total observations and I is the number of groups

From eq (2) - eq (1),

$$(n - 1) - (n - I) = 31 - 24$$

$$n - 1 - n + I = 7$$

$$I - 1 = 7$$

$$I = 8$$

Therefore, the number of groups are (I): 8.

(b) (4 points) Fill in the rest of the table. Values to be calculated are indicated by a “.” Please show how you compute the values for your calculations.

```
SST <- 70907
SSW <- 35088
SSB <- SST - SSW
SSB
```

```
## [1] 35819
```

```
I <- 8
dfW <- 24
dfT <- 31
dfB <- I - 1
dfB
```

```
## [1] 7
```

```
MSW <- SSW / dfW
MSW
```

```
## [1] 1462
```

```
MSB <- SSB / dfB
MSB
```

```
## [1] 5117
```

```
F_statistic <- MSB / MSW
F_statistic
```

```
## [1] 3.5
```

```
p_value <- 1-pf(F_statistic, df1=dfB, df2=dfW)
p_value
```

```
## [1] 0.009941808
```

Replacing values in the table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic	p-value
Between Groups	35819	7	5117	3.5	0.009941808
Within Groups	35088	24	1462		
Total	70907	31			

(c) (2 points) What is your conclusion from the one-way ANOVA analysis? State the hypothesis you are testing and what your decision/strength of evidence are.

Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \dots = \mu_I$. All the population means are equal

Alternative Hypothesis: $H_A : \mu_i \neq \mu_j$ At least two population means are different.

Conclusion:

The value of the test statistic is $F - statistic = 3.5$ and the $p - value = 0.009941808$. We reject the null hypothesis $H_0 : \mu_1 = \mu_2 = \dots = \mu_I$ since the obtained $p - value = 0.009941808$ is much lesser than the significance level ($\alpha = 0.05$) (assumption), in favor of alternative hypothesis. So basically $p - value$ is close to 0 and we have a strong evidence that we reject the null hypothesis and can conclude that at least two population means are different.

Question 2 (9 points) - Modified from *Sleuth* 5.25

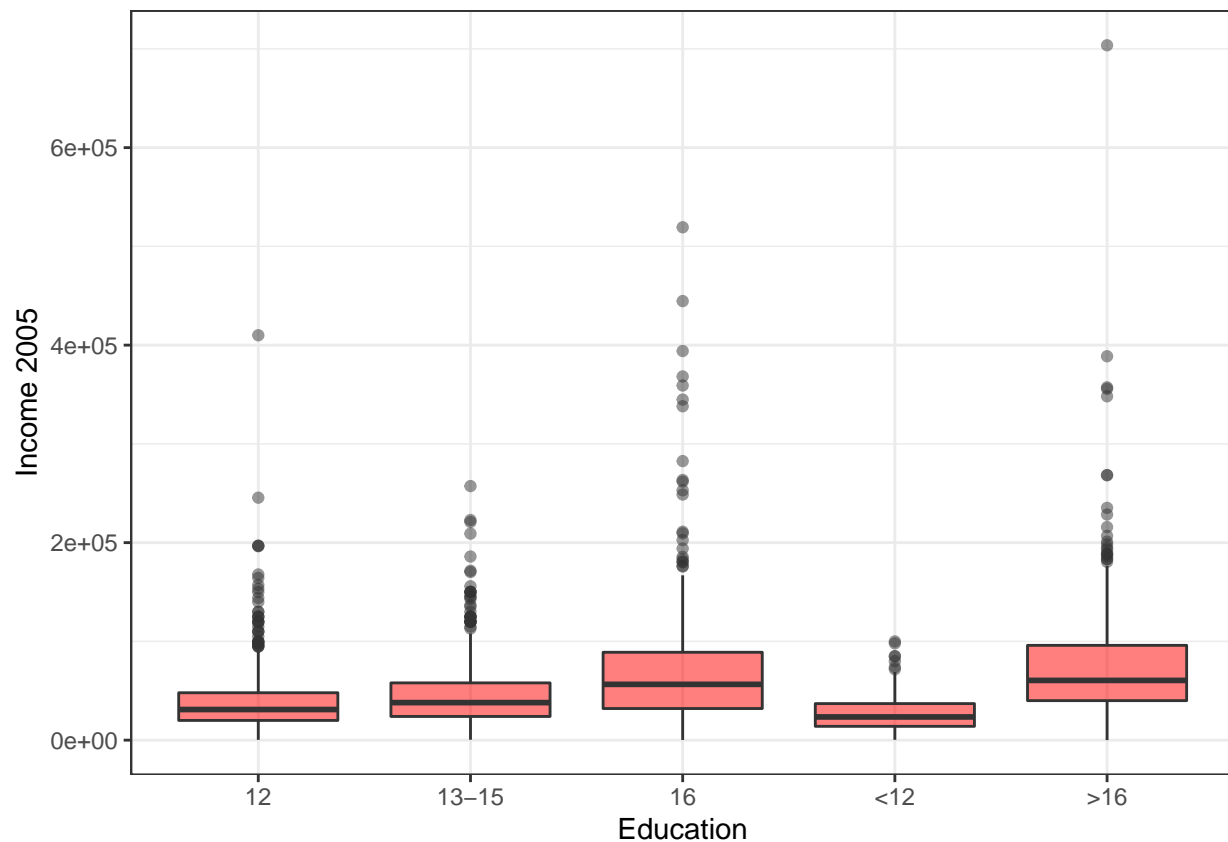
The data file `ex0525` contains annual incomes in 2005 of a random sample of 2,584 Americans who were selected for the National Longitudinal Survey of Youth in 1979 and who had paying jobs in 2005. The data set also includes a code for the number of years of education that each individual had completed by 2006: `<12`, `12`, `13-15`, `16`, and `>16`. Perform an analysis of variance *by hand* (i.e. not using the built-in anova functions like `lm()` and `anova()`) to assess whether or not the population mean 2005 incomes were the same in all five education groups. Work through the following steps:

(a) (1 point) Create a side-by-side boxplot of 2005 income grouped by education category.

```
data(ex0525)
names(ex0525)
```

```
## [1] "Subject"    "Educ"       "Income2005"
```

```
ggplot(ex0525, aes(x=Educ, y=Income2005)) +
  geom_boxplot(fill = "red", alpha = 0.5) +
  labs(y = "Income 2005", x = "Education") +
  theme_bw()
```



(b) (2 points) Find the grand mean and the mean of each of the five education groups.

```
grandmean <- mean(ex0525$Income2005)
grandmean
```

```
## [1] 49417
```

```
# <12, 12, 13-15, 16, and >16
mean_less_12 <- mean(ex0525$Income2005[ex0525$Educ == "<12"])
mean_less_12
```

```
## [1] 28301.45
```

```
mean_12 <- mean(ex0525$Income2005[ex0525$Educ == "12"])
mean_12
```

```
## [1] 36864.9
```

```
mean_13_15 <- mean(ex0525$Income2005[ex0525$Educ == "13-15"])
mean_13_15
```

```
## [1] 44875.96
```

```
mean_16 <- mean(ex0525$Income2005[ex0525$Educ == "16"])
mean_16
```

```
## [1] 69996.97
```

```
mean_greater_16 <- mean(ex0525$Income2005[ex0525$Educ == ">16"])
mean_greater_16
```

```
## [1] 76855.46
```

(c) (2 points) Find the sums of squares between and within groups.

```
# Create a vector by group
groupmeans <- c(mean_less_12, mean_12, mean_13_15, mean_16, mean_greater_16)
```

```
# sums of squares within group
SSW_2 <- sum((ex0525$Income2005[ex0525$Educ == "<12"]-groupmeans[1])^2) +
  sum((ex0525$Income2005[ex0525$Educ == "12"]-groupmeans[2])^2) +
  sum((ex0525$Income2005[ex0525$Educ == "13-15"]-groupmeans[3])^2) +
  sum((ex0525$Income2005[ex0525$Educ == "16"]-groupmeans[4])^2) +
  sum((ex0525$Income2005[ex0525$Educ == ">16"]-groupmeans[5])^2)
SSW_2
```

```
## [1] 4.951743e+12
```

```
# sums of squares between group = total - within
SST_2 <- sum((ex0525$Income2005 - grandmean)^2)
SST_2
```

```
## [1] 5.639978e+12
```

```
SSB_2 <- SST_2 - SSW_2
SSB_2
```

```
## [1] 688235137516
```

(d) (1 point) Find the mean squares between and within groups.

```
n_2 <- nrow(ex0525)
I_2 <- 5
dfW_2 <- n_2 - I_2
dfB_2 <- I_2 - 1

MSW_2 <- SSW_2 / dfW_2
MSW_2
```

```
## [1] 1920024320
```

```
MSB_2 <- SSB_2 / dfB_2
MSB_2
```

```
## [1] 172058784379
```

(e) (1 point) Find the F -statistic and p -value.

```
F_statistic_2 <- MSB_2 / MSW_2
F_statistic_2
```

```
## [1] 89.61282
```

```
p_value_2 <- 1-pf(F_statistic_2, df1=dfB_2, df2=dfW_2)
p_value_2
```

```
## [1] 0
```

(f) (1 point) State the conclusion of your test.

Null Hypothesis: $H_0 : \mu_{>12} = \mu_{12} = \mu_{13-15} = \mu_{16} = \mu_{>16}$. All the population means (Edu with age: <12, 12, 13-15, 16, and >16) are equal.

Alternative Hypothesis: $H_A : \mu_i \neq \mu_j$ At least two population means are different where $i, j \in <12, 12, 13-15, 16, >16$ and $i \neq j$.

The test statistic value $F_{statistic} = 89.61282$ and p -value is very, very small – so close to zero that 0 gets printed out due to rounding. We would reject the null hypothesis that the population mean Income2005 are all the same in favor of alternative hypothesis. It appears that at least one of the population mean Income2005 is different from the others.

(g) (1 point) We can also state things we have calculated in the “model testing/comparison” framework (You should not need to calculate anything new for this part). What is the extra sum of squares? What is the pooled variance?

Extra sum of squares = Sum of squares (Reduced) - Sum of squares (Full) = SST - SSW = Sum of squares between (SSB)

Therefore, the extra sum of squares is equal to 688235137516.

The mean sum of squares (Full) = MSW is same as pooled variance.

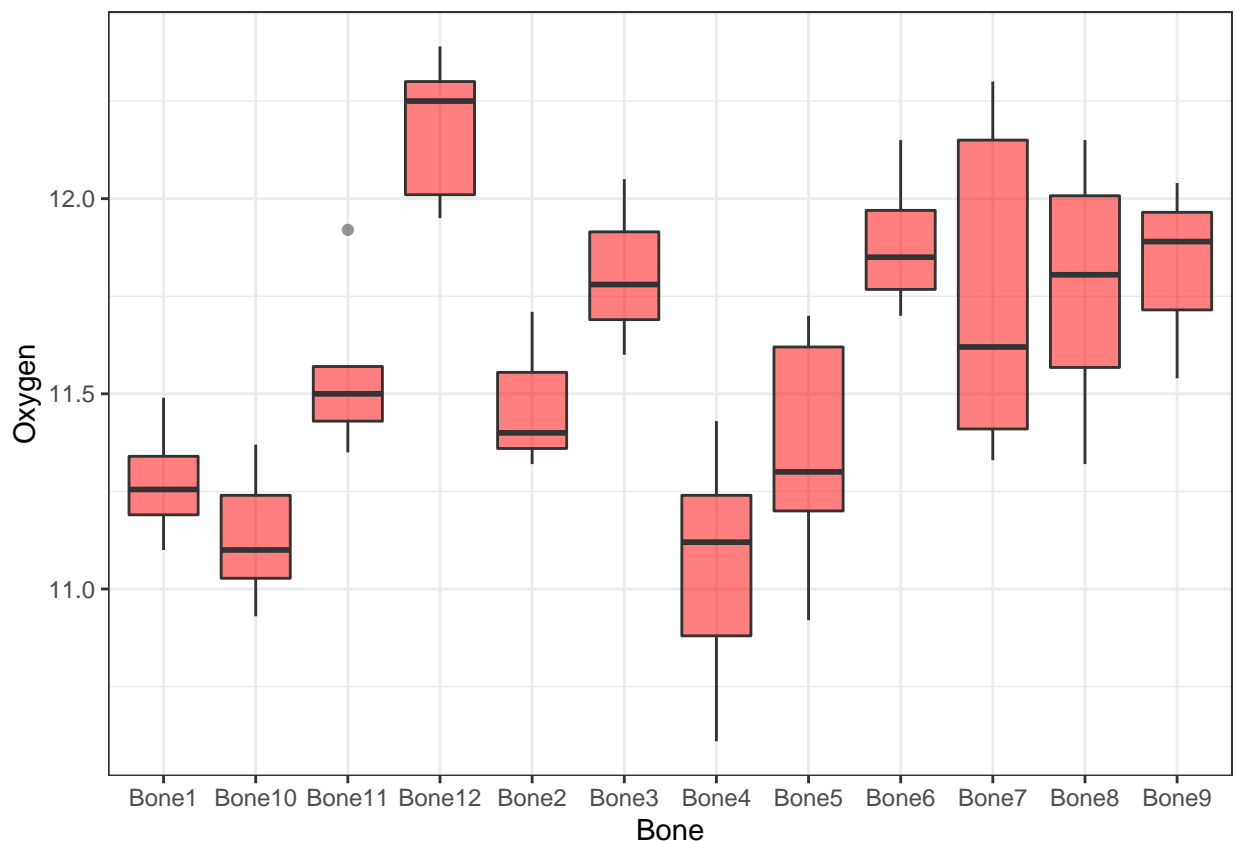
Therefore, the pooled variance is equal to 1920024320.

Question 3 (9 points) - from *Sleuth* 5.23

Was *Tyrannosaurus Rex* warm-blooded? Several measurements of the oxygen isotopic composition of bone phosphate in each of 12 bone specimens from a single *Tyrannosaurus rex* skeleton were taken. It is known that the oxygen isotopic composition of vertebrate bone phosphate is related to the body temperature at which the bone forms. Differences in means at different bone sites would indicate non-constant temperatures throughout the body. Minor temperature differences would be expected in warm-blooded animals. Is there evidence that the means are different for the different bones? The data are in `ex0523` in the `Sleuth3` library.

(a) (2 points) Plot the oxygen isotopic composition for each of the bones using a side-by-side boxplot. Comment on whether or not you think the population means are the same for all 12 bones based on your plot.

```
data(ex0523)
ggplot(ex0523, aes(x=Bone, y=Oxygen)) +
  geom_boxplot(fill = "red", alpha = 0.5) +
  labs(y = "Oxygen", x = "Bone") +
  theme_bw()
```



From the above boxplot, it is clear that the population means are not the same for all 12 bones since the means of each of the 12 bones are completely different from one another i.e., Bone1 is completely different from (Bone 6,7,8,9,12), Bone12 is completely different from the rest of the Bones and similarly, for others. Therefore, all the population mean oxygen isotopic compositions are not the same in the 12 bone types.

(b) (2 points) Perform an analysis of variance to test whether or not all the population mean oxygen isotopic compositions are the same in the 12 bone types. State your p -value and conclusion of the test. You may use the built-in ANOVA functions in R.

```
anova(lm(Oxygen ~ Bone, data=ex0523))
```

```
## Analysis of Variance Table
##
## Response: Oxygen
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Bone       11  6.0675  0.55159    7.4268 9.73e-07 ***
## Residuals  40  2.9708  0.07427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null hypothesis: $H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9 = \mu_{10} = \mu_{11} = \mu_{12}$. All the population mean oxygen isotopic compositions are the same in the 12 bone types

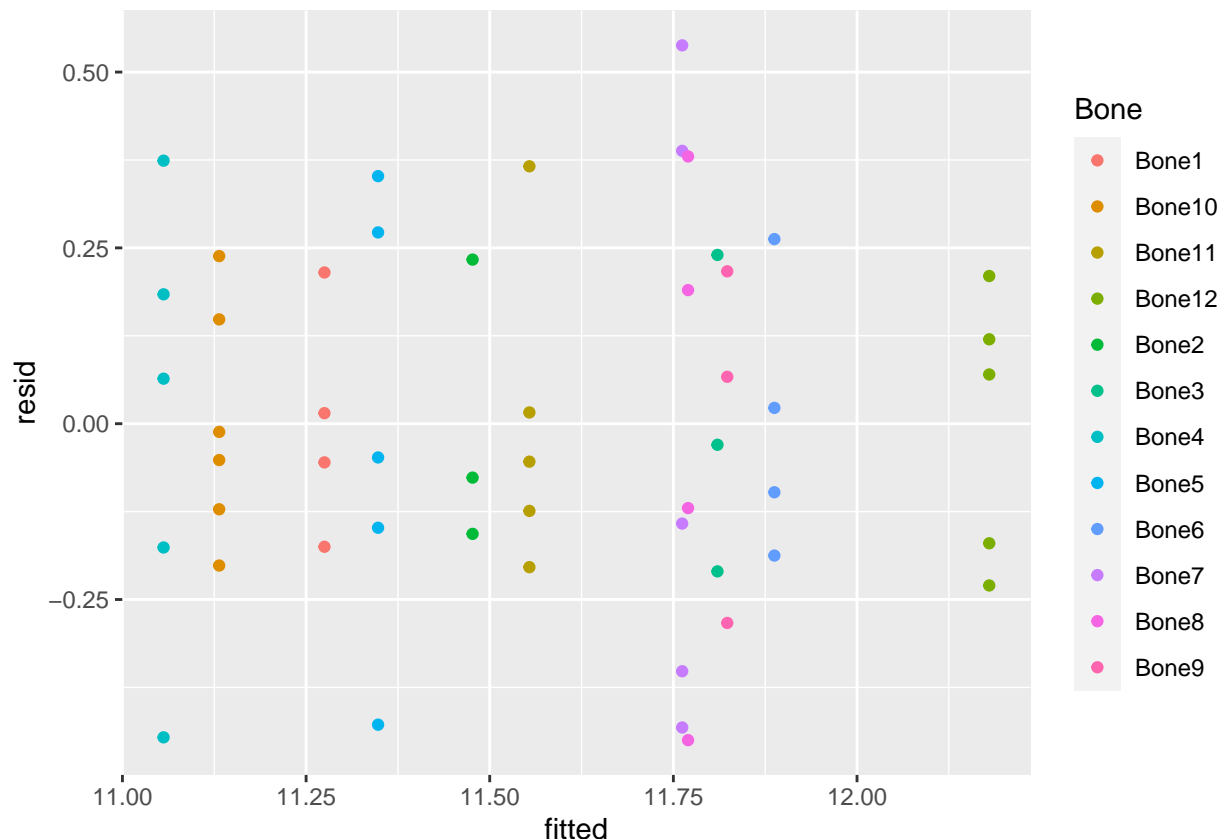
Alternative hypothesis: $H_A : \mu_i = \mu_j$ where $i, j \in [1, 12]$ and $i \neq j$. At least two of the population mean oxygen isotopic compositions in the 12 bone types are not the same.

Conclusion:

The test statistic F -value = 7.4268 and p -value = $9.73e - 07$. We reject the null hypothesis that all the population mean oxygen isotopic compositions are the same in the 12 bone types since the p -value = $9.73e - 07$ is much less than the significance level (0.05) and is nearly close to 0. Hence, we have sufficient evidence to conclude that the population mean oxygen isotopic compositions are not the same in the 12 bone types and there can be at least two of the population mean oxygen isotopic compositions in the 12 bone types that are not the same.

(c) (2 points) Assess the assumption that the population variances are the same in each group by creating a diagnostic plot using the residuals (See Lecture 18 for help with this). Does this assumption appear to have been met?

```
mod <- lm(Oxygen ~ Bone, data=ex0523)
ex0523$fitted <- mod$fitted
ex0523$resid <- mod$resid
ggplot(ex0523, aes(x=fitted, y=resid, color=Bone)) + geom_point()
```



Based on the above diagnostic plot, it is obvious that the equal population variance assumption doesn't met. We can notice that the variability for some groups is much larger/smaller than others i.e., variances of Bone7 is much larger than Bone1.

(d) (3 points) Perform a Kruskal-Wallis test using the `kruskal.test()` function. What do you conclude from this test? Compare your conclusion with your result from the analysis of variance in part (b).

```
kruskal.test(Oxygen ~ Bone, data=ex0523)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Oxygen by Bone
## Kruskal-Wallis chi-squared = 34.938, df = 11, p-value = 0.0002537
```

Null hypothesis: $H_0 : M_i = M_j$ where $i, j \in [1, 12]$ and $i \neq j$. The population centers oxygen isotopic compositions are the same in the 12 bone types

Alternative hypothesis: $H_A : M_i \neq M_j$ where $i, j \in [1, 12]$ and $i \neq j$. The population centers oxygen isotopic compositions centers of at least one of the groups is different.

Conclusion: The p -value = 0.0002537. We reject the null hypothesis that all the population centers oxygen isotopic compositions are the same in the 12 bone types since the p -value = 0.0002537 is much less than

the significance level (0.05) and is nearly close to 0. Hence, we have sufficient evidence to conclude that the population centers of oxygen isotopic compositions are not the same in the 12 bone types.

Both the tests came to similar conclusions of rejecting the null hypothesis. However, if residual plots indicate a problem with the ANOVA assumptions (Or, if sample sizes are very small and normality is suspect), then we can use Kruskal-Wallis Nonparametric Analysis of Variance as an alternative to ANOVA where Kruskal-Wallis tests uses centers to determine the results.

Question 4 (4 points)

(a) (2 points) In comparing 10 groups, a researcher notices that the sample mean of group 7 is the largest and the sample mean of group 3 is the smallest. The researcher then decides to test the hypothesis that $\mu_7 - \mu_3 = 0$. Why should a multiple comparison procedure be used even though there is only one comparison being made?

The researchers compared 10 groups and found that the sample mean of group 7 is largest and the sample mean of group 3 is the smallest. However, if researchers want to test the hypothesis that $\mu_7 - \mu_3 = 0$, they cannot simply do one comparison. There's a chance that the obtained results are not be significant since we have only one comparison where 3 and 7 may not be the smallest and largest, respectively.

Therefore, performing multiple comparison would give us more comparisons (with all the groups) and provides evidence of whether the sample mean of group 7 is the largest and the sample mean of group 3 is the smallest of all the 10 groups.

Hence, researchers need to do a multiple comparison procedure even though there is only one comparison being made to get significant results and avoid Type I errors.

b) (2 points) When choosing coefficients for a contrast, does the choice of $\{C_1, C_2, \dots, C_I\}$ give a different t -statistic than the choice of $\{4C_1, 4C_2, \dots, 4C_I\}$? Explain why or why not.

The change in the coefficients directly affects the value of γ since it is dependent on the coefficients. If $\{C_1, C_2, \dots, C_I\}$ gives the γ a value of 1x then $\{4C_1, 4C_2, \dots, 4C_I\}$ gives γ a value of 4x i.e., 4 times the original. In addition, the Standard Error also changes since the variance is affected by the multiple of 4. Therefore, equivalent changes in the γ (g) and Standard Error makes the overall change negligible. In other words, the multiple of 4 gets cancelled since it is common in both γ (g) and Standard Error

Hence, the choice of $\{C_1, C_2, \dots, C_I\}$ gives a same t -statistic as the choice of $\{4C_1, 4C_2, \dots, 4C_I\}$.

Question 5 (9 points) - Modified from *Sleuth* 6.17

The relative head length (RHL) is measured for adders (a type of snake) on the Swedish mainland and on groups of islands in the Baltic Sea. Relative head length is adjusted for overall body length, determined separately for males and females. The data are below and additionally you know that the pooled estimate of standard deviation of the RHL measurements was 11.72 based on 230 degrees of freedom.

```
adder <- data.frame(Locality = c("Uppsala", "In-Fredeln", "Inre Hamnnskar", "Norrpada",  
                                "Karringboskar", "Angskar", "SvenskaHagarna"),  
                   SampleSize = c(21, 34, 20, 25, 7, 82, 48),  
                   meanRHL = c(-6.98, -4.24, -2.79, 2.22, 1.27, 1.88, 4.98))
```

Consider the question: “Is the *average of the mean relative head lengths for snakes on the Swedish mainland* equal to the *average of the mean relative head lengths for snakes on islands in the Baltic Sea*?” Uppsala is the mainland, and the other six localities refer to islands in the Baltic Sea.

(a) (3 points) Write out the coefficients for the linear combination you would use to test this question, and state the null hypothesis you would be testing using statistical notation.

The coefficients for the linear combination are:

$C1 = 1$ $C2 = -(1/6)$, $C3 = -(1/6)$, $C4 = -(1/6)$, $C5 = -(1/6)$, $C6 = -(1/6)$, $C7 = -(1/6)$.

Null Hypothesis: $H_0 : 1\mu_1 - (1/6) * (\mu_2 + \mu_3 + \mu_4 + \mu_5 + \mu_6 + \mu_7) = 0$ (equivalent to $H_0 : \gamma = 0$) The *average of the mean relative head lengths for snakes on the Swedish mainland* equal to the *average of the mean relative head lengths for snakes on islands in the Baltic Sea*

Alternative Hypothesis: $H_A : 1\mu_1 - (1/6) * (\mu_2 + \mu_3 + \mu_4 + \mu_5 + \mu_6 + \mu_7) \neq 0$ (equivalent to $H_0 : \gamma \neq 0$). The *average of the mean relative head lengths for snakes on the Swedish mainland* is not equal to the *average of the mean relative head lengths for snakes on islands in the Baltic Sea*

(b) (4 points) What is the t -statistic for testing the hypothesis in part (a)? Please include in your answer your computed values of g and the standard error of g . Compute these values by hand.

```
# meanRHL = c(-6.98, -4.24, -2.79, 2.22, 1.27, 1.88, 4.98)
```

```
m1 <- -6.98
```

```
m2 <- -4.24
```

```
m3 <- -2.79
```

```
m4 <- 2.22
```

```
m5 <- 1.27
```

```
m6 <- 1.88
```

```
m7 <- 4.98
```

```
g <- 1*m1 - (1/6)*(m2+m3+m4+m5+m6+m7)
```

```
g
```

```
## [1] -7.533333
```

```

# SampleSize = c(21, 34, 20, 25, 7, 82, 48)
s1 <- 21
s2 <- 34
s3 <- 20
s4 <- 25
s5 <- 7
s6 <- 82
s7 <- 48

# Given, pooled estimate of standard deviation of the RHL measurements
# was 11.72 based on 230 degrees of freedom
sp2 <- 11.72 ^ 2
mainland <- (1*1)/(s1)
baltic_sea <- ((-1/6)^2) * ((1/s2) + (1/s3) + (1/s4) + (1/s5) + (1/s6) + (1/s7))

SE <- sqrt(sp2 * (mainland + baltic_sea))
SE

## [1] 2.769041

t_statistic <- (g - 0) / SE
t_statistic

```

```
## [1] -2.720557
```

(c) (2 points) Find the resulting p -value and state your conclusion (Be sure to write a *complete* conclusion for this test including all the relevant details and interpretations).

```

# Given, 230 degrees of freedom
p_value_5 <- 2 * (1 - pt(abs(-2.720557), df = 230))
p_value_5

```

```
## [1] 0.007015258
```

The t-statistic value is -2.720557 and p -value = 0.007015258. We reject the null hypothesis since the p -value = 0.007015258 is less than the significance level (0.05 and 0.01), in favor of alternative hypothesis. Therefore, we have sufficient evidence to conclude that the *average of the mean relative head lengths for snakes on the Swedish mainland* is not equal to the *average of the mean relative head lengths for snakes on islands in the Baltic Sea*.

Question 6 (12 points) - Modified from *Sleuth* 6.21

Reconsider the education and future income data from your last homework (data: `ex0525`). Find p -values and 95% confidence intervals for the difference in means for all pairs of education groups in the following ways:

(a) (2 points) Using the Tukey-Kramer procedure.

```
data(ex0525)

Education_Mod <- lm(Income2005 ~ Educ, data = ex0525)

Education_Tukey <- glht(Education_Mod, linfct = mcp(Educ = "Tukey"))

summary(Education_Tukey)
```

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Linear Hypotheses:
##           Estimate Std. Error t value Pr(>|t|)
## 13-15 - 12 == 0      8011      2201   3.639 0.002460 **
## 16 - 12 == 0      33132      2571  12.885 < 1e-04 ***
## <12 - 12 == 0     -8563      4000  -2.141 0.192860
## >16 - 12 == 0     39991      2649  15.098 < 1e-04 ***
## 16 - 13-15 == 0    25121      2774   9.058 < 1e-04 ***
## <12 - 13-15 == 0   -16574      4133  -4.010 0.000541 ***
## >16 - 13-15 == 0    31980      2846  11.239 < 1e-04 ***
## <12 - 16 == 0     -41696      4341  -9.604 < 1e-04 ***
## >16 - 16 == 0       6858      3140   2.184 0.176398
## >16 - <12 == 0    48554      4388  11.066 < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
confint(Education_Tukey)

##
## Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Quantile = 2.705
```

```
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##      Estimate      lwr      upr
## 13-15 - 12 == 0    8011.0607  2056.7376 13965.3838
## 16 - 12 == 0      33132.0768 26176.7327 40087.4210
## <12 - 12 == 0     -8563.4475 -19383.5492  2256.6541
## >16 - 12 == 0     39990.5665 32825.5456 47155.5874
## 16 - 13-15 == 0   25121.0161 17618.7624 32623.2699
## <12 - 13-15 == 0 -16574.5083 -27754.0252 -5394.9913
## >16 - 13-15 == 0  31979.5058 24282.4593 39676.5522
## <12 - 16 == 0    -41695.5244 -53438.7968 -29952.2520
## >16 - 16 == 0      6858.4897 -1636.6485 15353.6278
## >16 - <12 == 0   48554.0140 36685.3511 60422.6770
```

(b) (2 points) Without adjusting for multiple comparisons.

```
# Fisher's LSD (unadjusted)
summary(Education_Tukey, test=adjusted("none"))
```

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Linear Hypotheses:
##      Estimate Std. Error t value Pr(>|t|)
## 13-15 - 12 == 0    8011      2201   3.639 0.000279 ***
## 16 - 12 == 0      33132     2571  12.885 < 2e-16 ***
## <12 - 12 == 0     -8563      4000  -2.141 0.032380 *
## >16 - 12 == 0     39991      2649  15.098 < 2e-16 ***
## 16 - 13-15 == 0   25121      2774   9.058 < 2e-16 ***
## <12 - 13-15 == 0 -16574      4133  -4.010 6.23e-05 ***
## >16 - 13-15 == 0  31980      2846  11.239 < 2e-16 ***
## <12 - 16 == 0    -41696      4341  -9.604 < 2e-16 ***
## >16 - 16 == 0      6858      3140   2.184 0.029062 *
## >16 - <12 == 0   48554      4388  11.066 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

```
confint(Education_Tukey, calpha = univariate_calpha())
```

```
##
## Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
```



```
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Quantile = 1.9609
## 95% confidence level
##
##
## Linear Hypotheses:
##           Estimate      lwr      upr
## 13-15 - 12 == 0    8011.0607  3694.7196 12327.4018
## 16 - 12 == 0      33132.0768 28090.0868 38174.0669
## <12 - 12 == 0     -8563.4475 -16407.0342  -719.8609
## >16 - 12 == 0     39990.5665  34796.5799 45184.5531
## 16 - 13-15 == 0   25121.0161 19682.5665 30559.4657
## <12 - 13-15 == 0 -16574.5083 -24678.6382  -8470.3783
## >16 - 13-15 == 0  31979.5058 26399.8492 37559.1623
## <12 - 16 == 0     -41695.5244 -50208.3257 -33182.7231
## >16 - 16 == 0       6858.4897   700.2894 13016.6899
## >16 - <12 == 0    48554.0140 39950.3161 57157.7120
```

(c) (3 points) What do you notice by comparing these two methods? Discuss the differences in which pairwise comparisons are *significant*, how wide the confidence intervals are, and which confidence intervals contain 0.

Differences in which pairwise comparisons are significant:

We can notice that if we did not use an adjustment procedure, Fisher's LSD (unadjusted), we would think all of the comparisons were significant at the 0.05 level since all the p -values are less than 0.05. But when we used the Tukey-Kramer adjustment, only eight of them were significant and remaining two are insignificant since their p -value is greater than 0.05.

Therefore, we Tukey-Kramer adjustment gives some good values which makes it less likely to reject the null hypothesis based on the obtained p values and has better pairwise comparisons than the Fisher's LSD (unadjusted).

How wide the confidence intervals are:

The confidence intervals are much wider in Tukey-Kramer procedure than that of Fisher's LSD (unadjusted) since the later procedure is done without adjustment. Compare the intervals of Tukey-Kramer with the Fisher's LSD (unadjusted) we notice that the point estimates are the same (as we expect), but the lower and upper bounds are changing.

For example, the lower-bounds and upper-bounds for the group (13-15 - 12 == 0) is (2057.1791, 13964.9423) in Tukey-Kramer whereas it is (3694.7196, 12327.4018) in Fisher's LSD (unadjusted) which shows us that the confidence intervals are wider in Tukey-Kramer procedure.

Which confidence intervals contain 0:

The groups (<12 - 12 == 0) and (>16 - 16 == 0) Tukey-Kramer procedure contains 0 in their confidence intervals. None of the groups from Fisher's LSD (unadjusted) contains 0 in their confidence intervals.

(d) (3 points) Use the Dunnett procedure to compare every other group to the group with 12 years of education. Look at both the p -values and confidence intervals. Which group means apparently differ from the mean for those with 12 years of education?

```
levels(ex0525$Educ) # print levels before re-leveling
```

```
## [1] "12"      "13-15" "16"      "<12"     ">16"
```

```
ex0525$Educ <- relevel(ex0525$Educ, ref = "12")
levels(ex0525$Educ) # print levels after re-leveling
```

```
## [1] "12"      "13-15" "16"      "<12"     ">16"
```

```
Education_Model <- lm(Income2005 ~ Educ, data = ex0525)
```

```
Education_Dunnett <- glht(Education_Model, linfct = mcp(Educ = "Dunnett"))
summary(Education_Dunnett)
```

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Linear Hypotheses:
##              Estimate Std. Error t value Pr(>|t|)
## 13-15 - 12 == 0      8011         2201   3.639  0.00111 **
## 16 - 12 == 0       33132         2571  12.885 < 1e-04 ***
## <12 - 12 == 0      -8563         4000  -2.141  0.11803
## >16 - 12 == 0      39991         2649  15.098 < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
confint(Education_Dunnett)
```

```
##
## Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Quantile = 2.4811
## 95% family-wise confidence level
##
##
```

```
## Linear Hypotheses:
##           Estimate      lwr      upr
## 13-15 - 12 == 0   8011.0607  2549.7136 13472.4078
## 16 - 12 == 0     33132.0768 26752.5862 39511.5675
## <12 - 12 == 0    -8563.4475 -18487.7210  1360.8259
## >16 - 12 == 0    39990.5665  33418.7588 46562.3742
```

The p -values for (13-15 - 12 == 0), (16 - 12 == 0), (>16 - 12 == 0) are all less than 0.05 significance level and their respective group means differ from the mean for those with 12 years of education since the p -value and confidence interval for the group <12 - 12 are 0.11804 (>0.05 significance) and 0 respectively (whereas the rest of the groups have p -value < 0.05 and confidence interval \neq 0).

Therefore, we have an evidence that groups have means are different (p -value < 0.05 and confidence interval \neq 0) with 12 years of education (p -values > 0.05, confidence interval contain 0). (Fail to reject null hypothesis: different means)

(e) (2 points) Taking all of these tests together, what general statements would you make about the relationship between Education and Income?

Taking all of these tests together, we can say that the population mean 2005 incomes were not the same in all five education groups since the p -value is less than 0.05 and confidence interval contains 0 in both the Tukey-Kramer procedure and Dunnett procedure. In general, we can say that the education group with age (<12-12) has a different mean from the rest of the groups and will have less income compared to that of the other groups (based on the CI and p -value).