FORE224/STAT202 Assignment 7 One-way analysis of variance (ANOVA) and post-hoc tests

Matthew Sherman

28th September, 2023

Question One

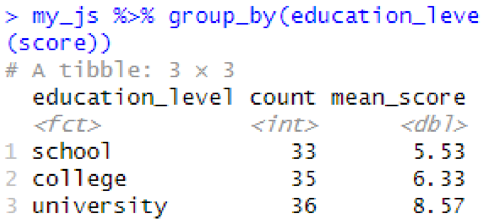
N/A

Question Two

```
library(tidyverse)
library(performance)
data <- read_csv("job_satisfaction1.csv") %>% mutate(education_level =
factor(education_level, levels = c("school", "college", "university")))
set.seed(16645573)
my_js <- data %>% sample_n(104)
```

Question Three

```
my_js %>% group_by(education_level) %>% summarise(count = n(), mean_score =
mean(score))
```



```
> my_js %>% group_by(education_level)
(score))
# A tibble: 3 x 3
  education_level count mean_score
  <fct>           <int>      <dbl>
1 school             33       5.53
2 college            35       6.33
3 university         36       8.57
```

Mean values:

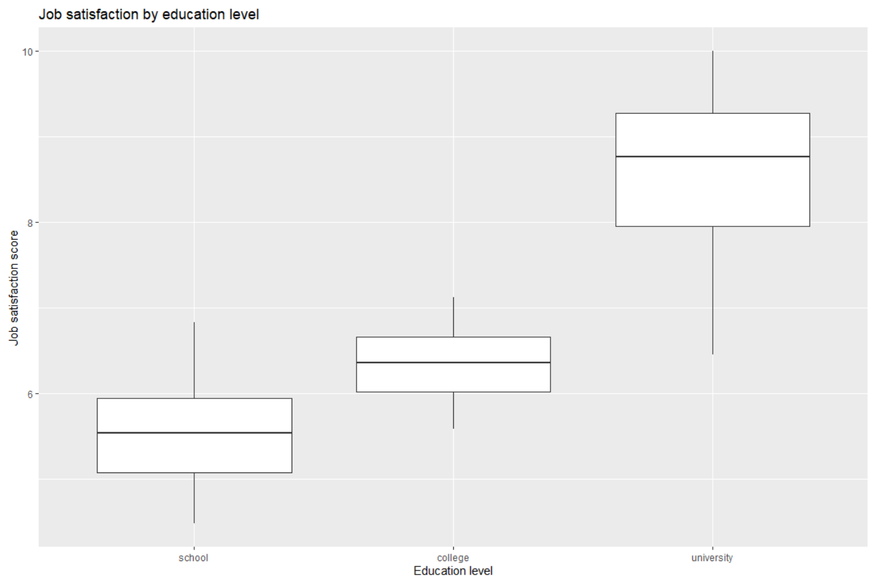
School – 5.53

College – 6.33

University – 8.57

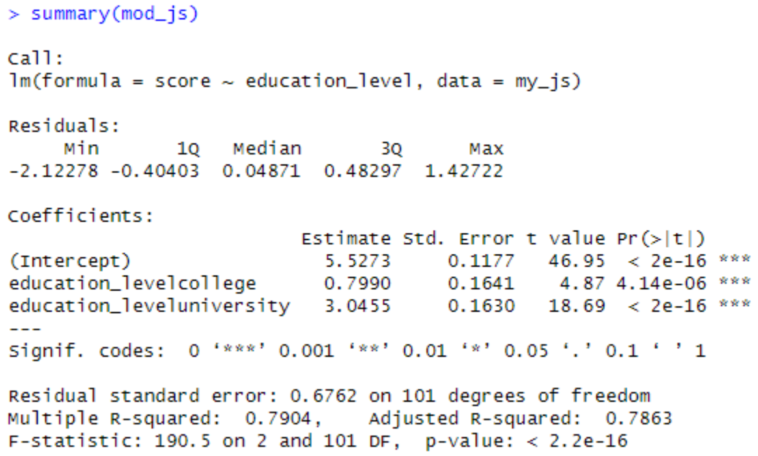
Question Four

```
my_js %>% ggplot(aes(x = education_level, y = score)) +
  geom_boxplot() +
  labs(
    title = "Job satisfaction by education level",
    x = "Education level",
    y = "Job satisfaction score"
  )
```



Question Five

```
mod_js <- lm(score ~ education_level, data = my_js)
summary(mod_js)
```



```
> summary(mod_js)

Call:
lm(formula = score ~ education_level, data = my_js)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.12278 -0.40403  0.04871  0.48297  1.42722 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 5.5273     0.1177   46.95  < 2e-16 ***
education_levelcollege      0.7990     0.1641    4.87 4.14e-06 ***
education_leveluniversity   3.0455     0.1630   18.69  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6762 on 101 degrees of freedom
Multiple R-squared:  0.7904,	Adjusted R-squared:  0.7863 
F-statistic: 190.5 on 2 and 101 DF,  p-value: < 2.2e-16
```

Explain in up to 80 words what the coefficient values in the model tell you and relate these coefficient values to the summary of the data you obtained in question 3.

The intercept coefficient value of 5.5273 represents the mean score for the school education level.

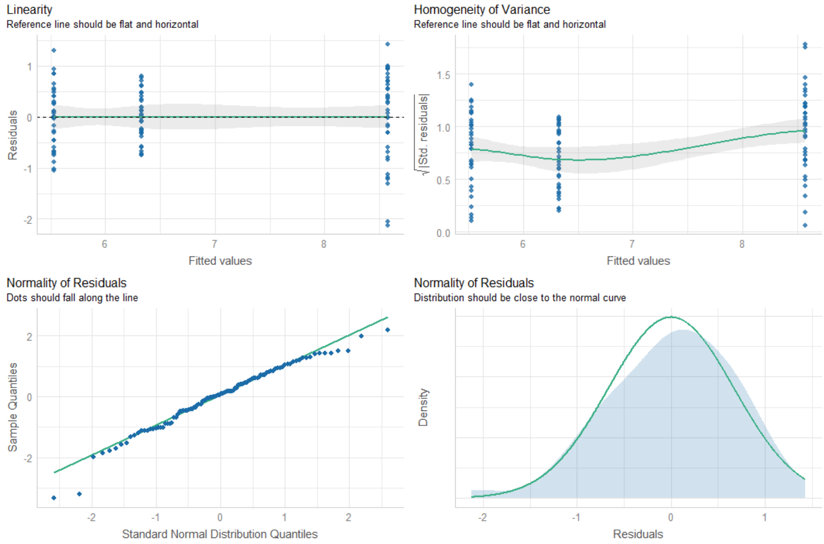
The education_levelcollege coefficient value of 0.7990 represents that on average, those who are college educated score 0.7990 higher than those who are school educated.

The education_leveluniversity coefficient value of 3.0455 represents that on average, those who are university educated score 3.0455 higher than those who are school educated.

These coefficient values relate to the summary of the data obtained in Q3 in that the Q3 summary shows higher mean score values the higher the education level.

Question Six

```
check_model(mod_js, check = c(
  "linearity",
  "homogeneity",
  "qq",
  "normality"
))
```



Explain in up to 100 words what assumptions the Homogeneity plot and Normal Q-Q plot give information about and whether these plots indicate anything unusual or wrong for this model.

The Normal Q-Q plot gives information about the assumption that residuals are normally distributed.

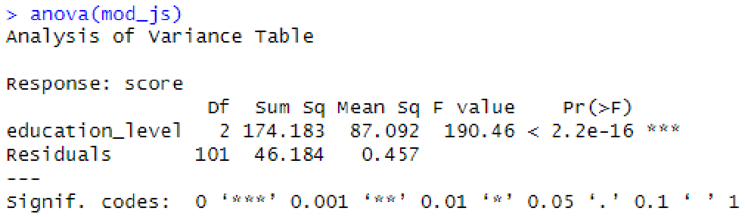
The Homogeneity plot gives information about the assumption that variances are the same for all populations.

The normality of residuals (Q-Q) plot shows the residuals falling roughly along the reference line with a small amount deviation from normality at each tail end. However, this deviation is not substantial and the assumption that residuals are normally distributed is still met.

The homogeneity of variance plot shows a reference line which is not flat or horizontal and instead is somewhat curved. The assumption that variances are the same for all populations is not met as there is not constant variance.

Question Seven

```
anova(mod_js)
```



```
> anova(mod_js)
Analysis of Variance Table

Response: score
                Df  Sum Sq Mean Sq F value    Pr(>F)    
education_level   2 174.183  87.092  190.46 < 2.2e-16 ***
Residuals       101  46.184   0.457                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

State in 30-50 words what null and alternative hypotheses the F-test statistic and its associated p-value shown in the anova output are testing, in the context of the variables used in this model.

The F-test statistic and its associated p-value in the ANOVA output test the null hypothesis that there is no significant difference in the score means for different education levels. They also test the alternative hypothesis that at least one pair of education levels has significantly different score mean values.

Give your conclusion based on this F-test and state what evidence that conclusion is based on (about 20-30 words in total).

I conclude that at least one pair of education levels have significantly different score mean values. This is because of the very low p-value (< 2.2e-16), so the null hypothesis is rejected.

Question Eight

How many pair-wise comparison tests would you need to make if you tested each group in the education_level factor compared to every other group?

Can use the formula M=(m*(m-1))/2 with m being the levels of a factor (which is 3 because there are the groups school, college, and university).

(3*(3-1))/2 = 6/2 = 3

Answer: 3 pair-wise comparison tests would need to be made

Show how the overall Type 1 error rate is calculated if a significance level of 0.05 is used for each test and give the result of this calculation (this is covered in Lecture 23). Answer in about 20 words in total.

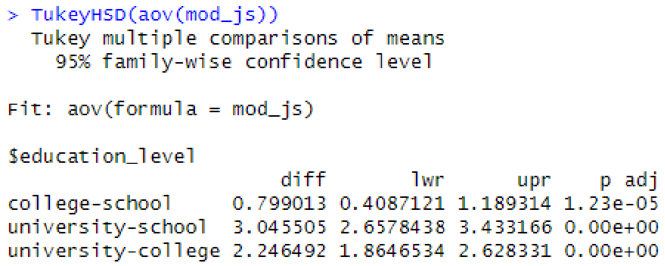
Can use the formula 1 - (1 - α)^M, with α being the significance level (0.05) and M being the number of pair-wise comparison tests (3).

1 - (1 – 0.05)^3 = 0.142625

Answer: The overall Type 1 error rate for this question is 0.142625

Question Nine

```
TukeyHSD(aov(mod_js))
```



```
> TukeyHSD(aov(mod_js))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mod_js)

$education_level
                       diff       lwr      upr    p adj
college-school     0.799013 0.4087121 1.189314 1.23e-05
university-school  3.045505 2.6578438 3.433166 0.00e+00
university-college 2.246492 1.8646534 2.628331 0.00e+00
```

Explain in about 50 words what the confidence intervals shown in the TukeyHSD function output are confidence intervals for.

They are adjusted pairwise confidence intervals for the differences in population means between the education level groups. The intervals give a range of values within which we can be 95% confident that the true population difference in means is in.

Explain in about 30-50 words whether the TukeyHSD confidence intervals show evidence of a difference between any of the education_level group population means and if so, which.

All of the pairwise functions (college-school, university-school, university-college) show confidence intervals which do not overlap, as well as all the p adj values for all being 0 or very close to 0. This indicates that there is statistically significant difference between all of the education_level group population means (college-school, university-school, university-college).

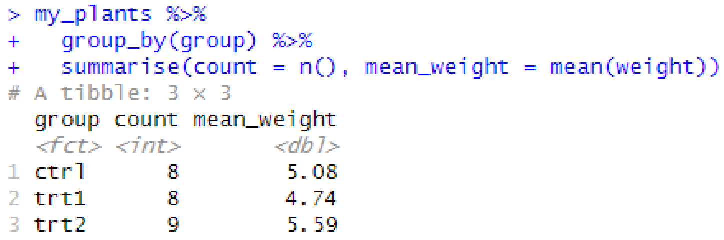
Question Ten

```
set.seed(16645573)
my_plants <- PlantGrowth %>% sample_n(25)
```

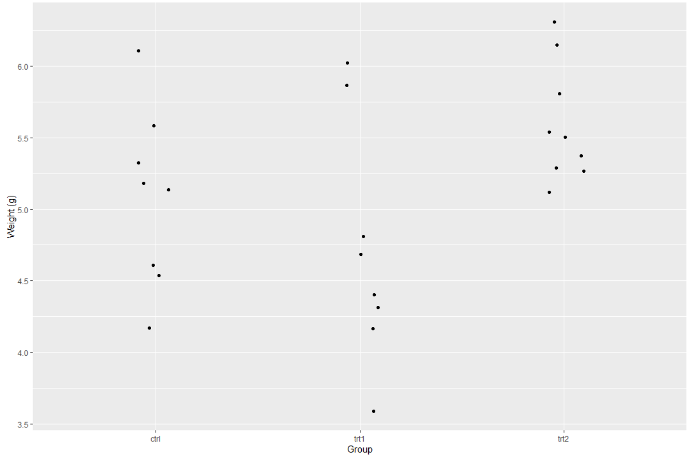
Question Eleven

```
my_plants %>%
  group_by(group) %>%
  summarise(count = n(), mean_weight = mean(weight))

my_plants_plot <- my_plants %>%
  ggplot(aes(x = group, y = weight)) +
  geom_jitter(width = 0.1) +
  labs(
    x = "Group",
    y = "Weight (g)"
  )
```



```
> my_plants %>%
+   group_by(group) %>%
+   summarise(count = n(), mean_weight = mean(weight))
# A tibble: 3 x 3
  group count mean_weight
  <fct> <int>       <dbl>
1 ctrl      8        5.08
2 trt1      8        4.74
3 trt2      9        5.59
```

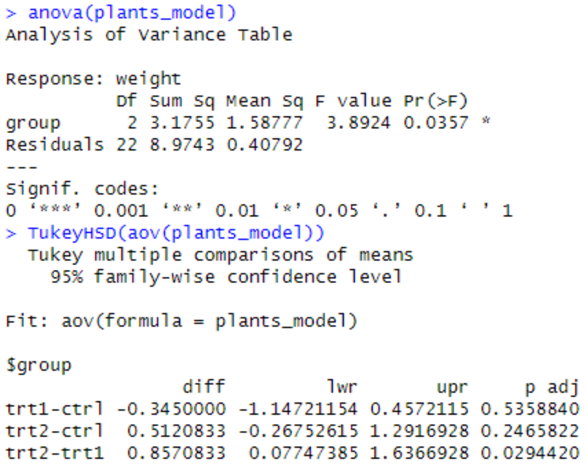


On the basis of this information, do you think that there might be differences in the population mean weights by group for the sampled populations? Summarise your thoughts in about 50-70 words in your report.

The summary shows the ctrl group having a mean weight of 5.08 grams, the trt1 group having a mean weight of 4.74 grams, and the trt2 group having a mean weight of 5.59 grams, so the mean weights are all different. Because of this, I think that there might be differences in the population mean weights by group for the sampled populations. However, other tests may need to be done to be more confident in this conclusion.

Question Twelve

```
plants_model <- lm(weight ~ group, data = my_plants)
anova(plants_model)
TukeyHSD(aov(plants_model))
```



```
> anova(plants_model)
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value Pr(>F)  
group      2 3.1755 1.58777  3.8924 0.0357 *
Residuals 22 8.9743 0.40792                 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(aov(plants_model))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = plants_model)

$group
                diff         lwr       upr     p adj
trt1-ctrl -0.3450000 -1.14721154 0.4572115 0.5358840
trt2-ctrl  0.5120833 -0.26752615 1.2916928 0.2465822
trt2-trt1  0.8570833  0.07747385 1.6366928 0.0294420
```

Comment in about 100 words on whether there is evidence of a difference between any of the pairs of group population mean weights and if so, which. You should explain clearly what values from the output you are using and how these justify your comments.

The ANOVA results show an F-value of 3.8924 with an associated p-value of 0.0357 for group. The p-value is less than the 0.05 significance level. This indicates that there is evidence for a difference in at least one pair of group population mean weights.

To check what exact pairs are different, the TukeyHSD test can be used.

The trt1-ctrl p adj value is 0.5358840, which is higher than the 0.05 significance level. Hence, there is no evidence of a difference of the population mean weights between these two groups.

The trt2-ctrl p adj value is 0.2465822, which is higher than the 0.05 significance level. Hence, there is no evidence of a difference of the population mean weights between these two groups.

The trt2-trt1 p adj value is 0.0294420, which is lower than the 0.05 significance level. Hence, there is evidence of a difference of the population mean weights between these two groups.

Full Code

```
library(tidyverse)
library(performance)
data <- read_csv("job_satisfaction1.csv") %>% mutate(education_level =
factor(education_level, levels = c("school", "college", "university")))
set.seed(16645573)
my_js <- data %>% sample_n(104)

my_js %>% group_by(education_level) %>% summarise(count = n(), mean_score =
mean(score))

my_js %>% ggplot(aes(x = education_level, y = score)) +
  geom_boxplot() +
  labs(
    title = "Job satisfaction by education level",
    x = "Education level",
    y = "Job satisfaction score"
  )

mod_js <- lm(score ~ education_level, data = my_js)
summary(mod_js)

check_model(mod_js, check = c(
  "linearity",
  "homogeneity",
  "qq",
  "normality"
))

check_homogeneity(mod_js)
check_normality(mod_js)

anova(mod_js)

TukeyHSD(aov(mod_js))

set.seed(16645573)
my_plants <- PlantGrowth %>% sample_n(25)

my_plants %>%
  group_by(group) %>%
  summarise(count = n(), mean_weight = mean(weight))
```

```
my_plants %>%
  ggplot(aes(x = group, y = weight)) +
  geom_jitter(width = 0.1) +
  labs(
    x = "Group",
    y = "Weight (g)"
  )

plants_model <- lm(weight ~ group, data = my_plants)
anova(plants_model)
TukeyHSD(aov(plants_model))
```