# Assignment7

## Question1

library(tidyverse)

jobs <- read\_csv('job\_satisfaction2.csv')

set.seed(17053777)

my\_js <- jobs %>%

sample\_n(104) %>%

mutate(education\_level = factor(education\_level,

levels = c("school", "college", "university"))) %>%

mutate(gender = factor(gender))

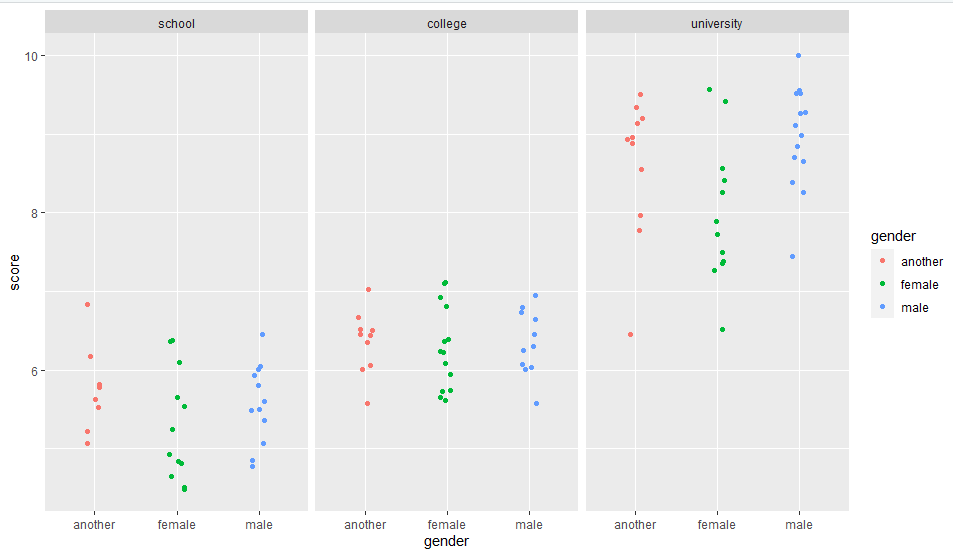
## Question2

my\_js %>% ggplot(aes( x = gender,

y = score, colour = gender )) +

geom\_jitter(width = 0.1) +

facet\_wrap(~education\_level) # does a nice layout :)



## Question3

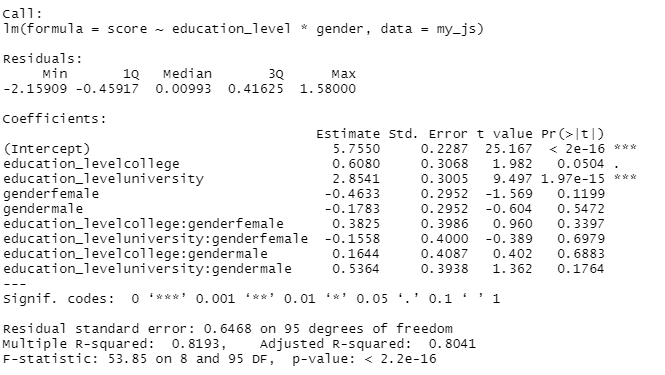
# If there is no interaction between gender and educational level, the difference in score of different gender is the same in different educational group.

# If there is interaction between educational level and gender, the mean score of different gender could be different in different educational level.

## Question4

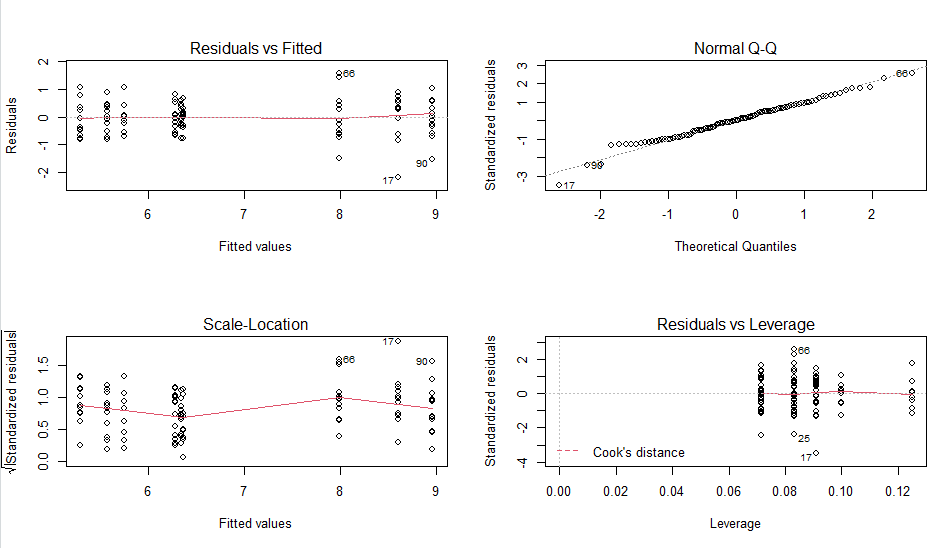
m1 <- lm(score ~ education\_level\*gender, data=my\_js)

summary(m1)



par(mfrow = c(2, 2))

plot(m1)

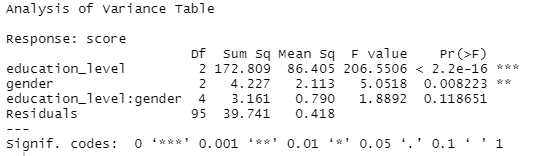
# From the graph, we can see there is nothing to worry about.

# From Normal Q-Q, we can see the data points fitted the line well, so we can say there are not too many deviations from the normality although there is an outlier around the red dash line on the tail.

# The scale-location plot also looks fairly close to constant.

## Question5

anova(m1)



# p-value for the F-test for the interaction between education level and gender is 0.118651,

# Null hypothesis is that there is no interaction between educational level and gender, p-value is greater than 0.05, so do not reject null hypothesis.

# so, we could drop the interaction in the model as there is no evidence of interaction.

## Question6

library(palmerpenguins)

male\_penguins <- penguins %>%

drop\_na() %>%

filter(sex == "male")

set.seed(17053777)

my\_pen\_m <- male\_penguins %>% sample\_n(150)

## Question7

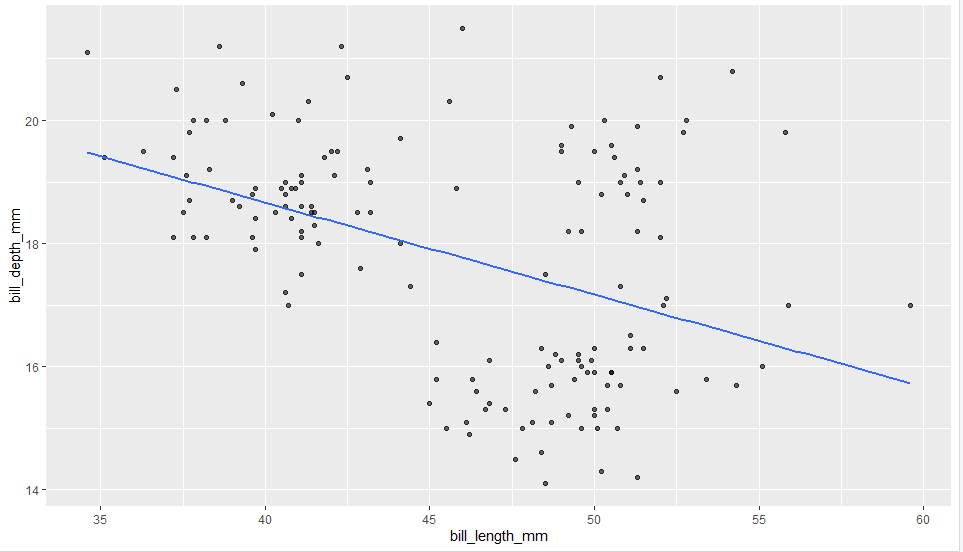
my\_pen\_m %>% ggplot(aes( x = bill\_length\_mm, y = bill\_depth\_mm )) +

geom\_point(alpha = 0.6) +

geom\_smooth(method = lm, se = FALSE)

# plot the single linear regression line

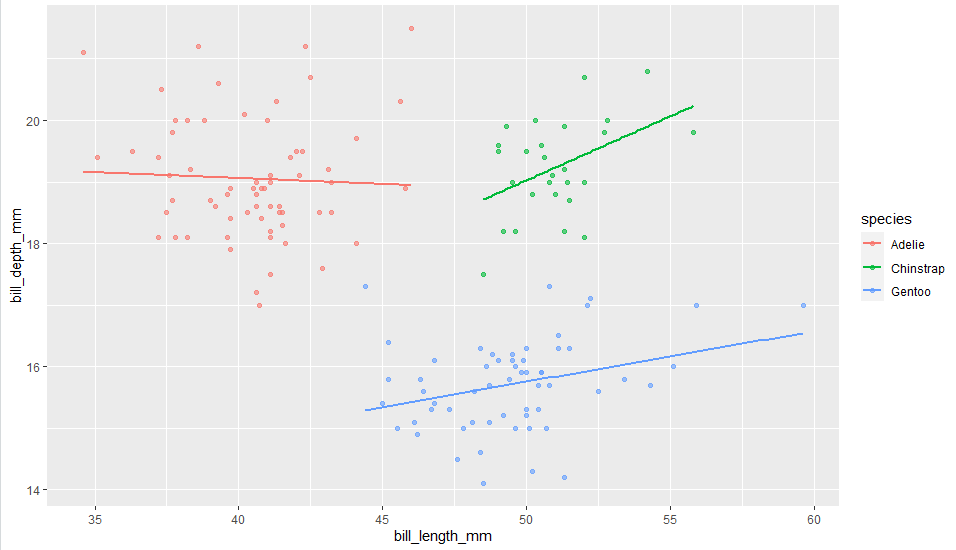
# second plot with coloured by species.



my\_pen\_m %>% ggplot(aes( x = bill\_length\_mm, y = bill\_depth\_mm, colour = species)) +

geom\_point(alpha = 0.6) +

geom\_smooth(method = lm, se = FALSE)



# with no species colour, there is one line

# but after adding the colour of species, there are 3 lines. the relationship between bill depth and bill length is different for 3 species.

# And we can see there are 3 different groups by different colour.

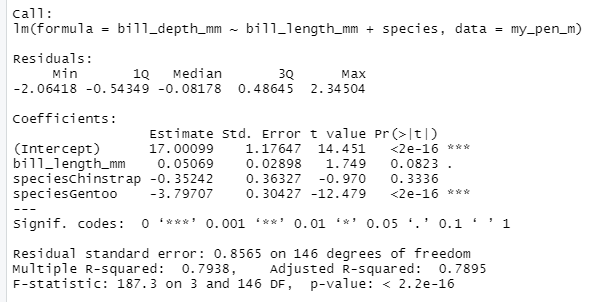
# so that we can avoid a biased assessment of the impact of bill length on bill depth.

## Question8

#no interaction

m2 <- lm(bill\_depth\_mm ~ bill\_length\_mm + species, data = my\_pen\_m)

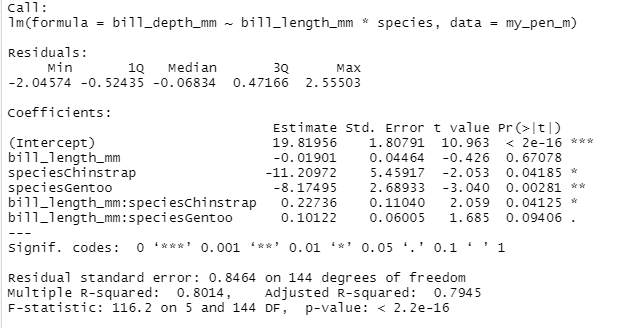
summary(m2)



#with interaction

m3 <- lm(bill\_depth\_mm ~ bill\_length\_mm \* species, data = my\_pen\_m)

summary(m3)



# For the model of no interaction, there is a common slope of three species.

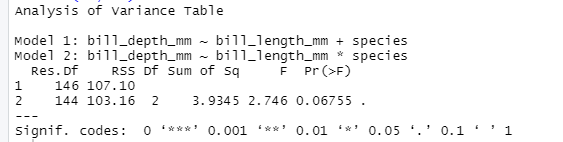
# for the model with interaction, there are three different slopes.

# from the graph of 3 different species, we can see only the slope of Adelie is negative, and close to 0 and the slope for other 2 are positive.

# But from the graph ignoring species, we can see the slope is negative.

## Question9

anova(m2, m3)



# the null hypothesis is that in the population model the coefficient of bill\_length:species interaction is 0.

# alternative hypotheses is that in the population model the coefficient of bill\_length:species interaction is not 0.

Question10

# adelie : y = 19.82 - 0.02 \* bill\_length

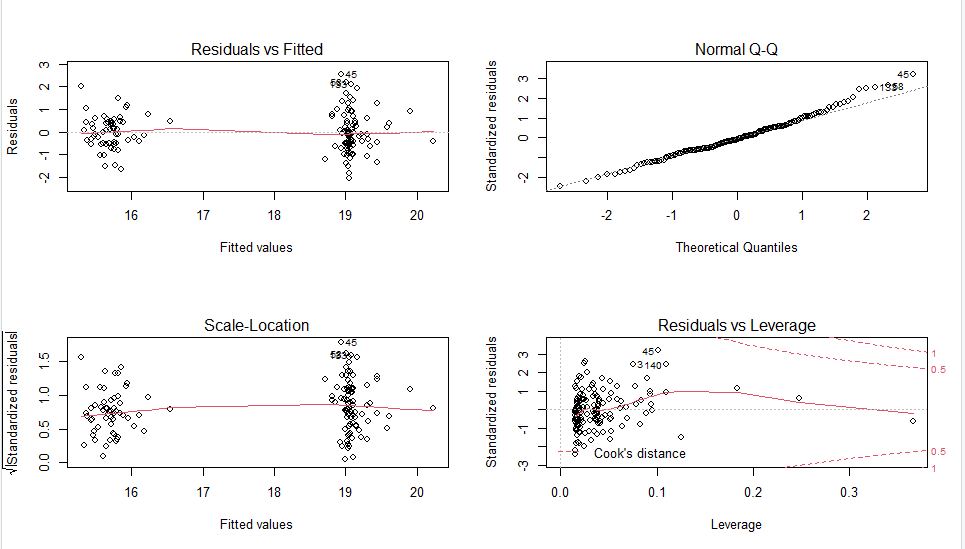
# chinstrap : y = 19.82 - 11.21 + ((- 0.02) + 0.23) \* bill\_length

# Gentoo : y = 19.82 - 8.17 + ((- 0.02) + 0.10) \* bill\_length

Question11

par(mfrow = c(2, 2))

plot(m3)



# the graph is fine overall

# From Normal Q-Q, we can see there is not too many deviations from the normality. All points fitting a normal distribution well here.

# From Residuals vs Leverage, we can see there is no outliers around 1 red dash line and 0.5 re line dash.