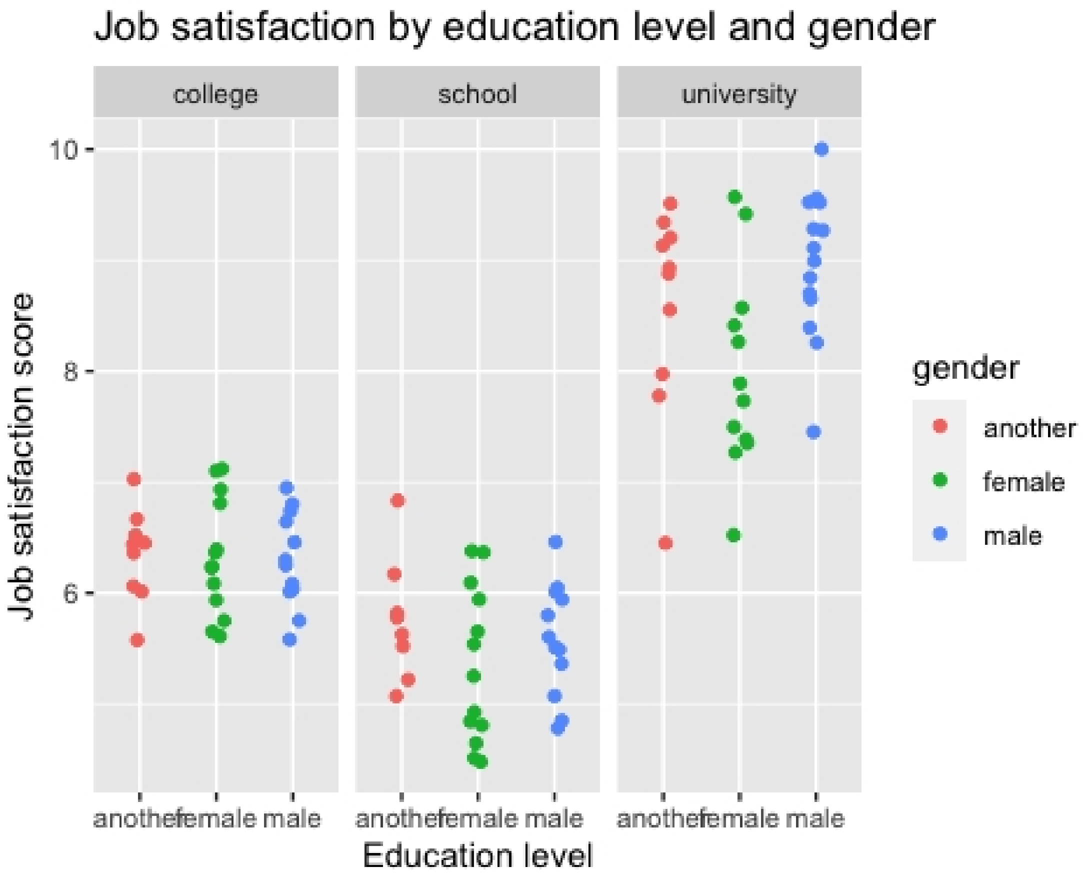
1. Load the tidyverse library.

```
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.1 ──

## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.5     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1

## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

js <- read_csv('job_satisfaction2.csv')

## Rows: 106 Columns: 3

## ── Column specification
────────────────────────────────────────────────────────

## Delimiter: ","
## chr (2): gender, education_level
## dbl (1): score

##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

set.seed(31878039)

my_js <- js %>%
  sample_n(104) %>%
  mutate(education_level <- factor(education_level,
 levels = c("school", "college", "university")),
 gender = factor(gender)
 )
```

2. Make a jittered scatter plot of score by education level and gender using ggplot.

```
my_js %>% ggplot(aes(
 x = gender, y = score,
 colour = gender
)) +
 geom_jitter(width = 0.1) +
 facet_wrap(~education_level) +
  labs(
 title = "Job satisfaction by education level and gender",
 x = "Education level",
 y = "Job satisfaction score"
 )
```

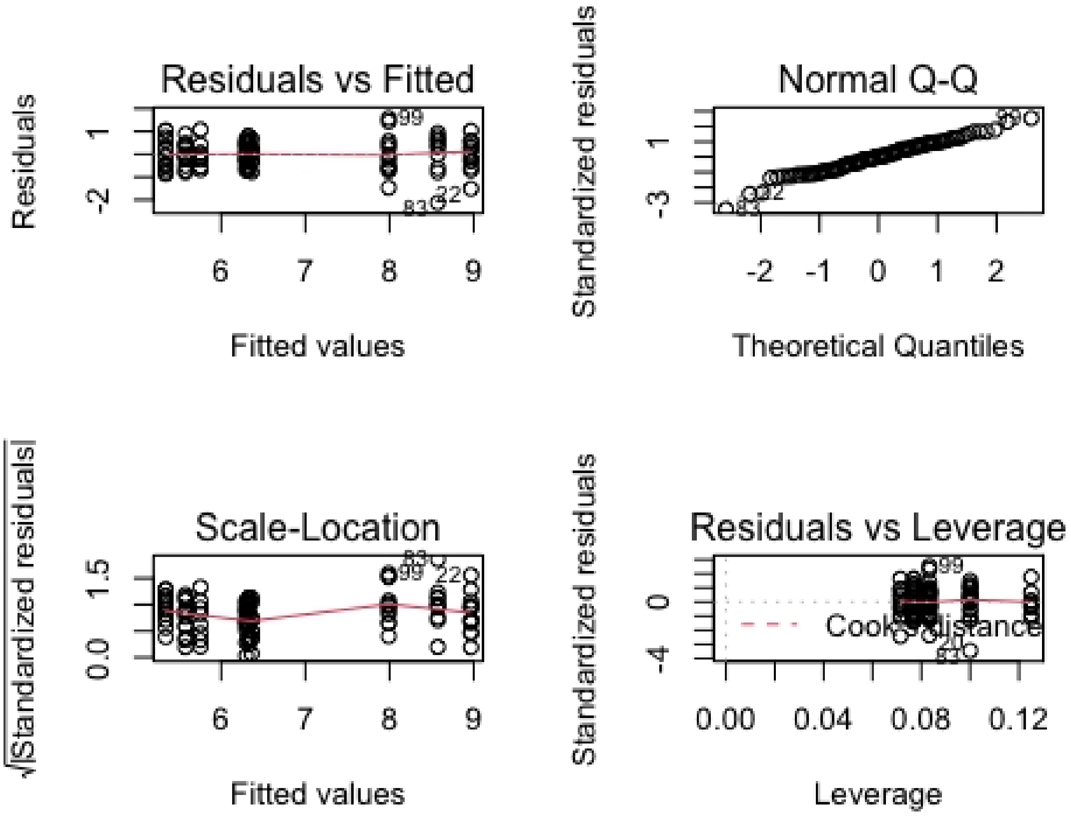


3. You are going to investigate whether there is evidence of an interaction between education_level and gender in the sampled populations. Explain briefly (in about 70 words) what it would mean for the relationship between population mean job satisfaction scores and education level and gender if there is such an interaction. You may find it helpful to use examples rather than a theoretical explanation.

If there is an interaction between education level and gender in predicting population mean job satisfaction scores, then gender and education level has an effect in the model. This means that there are differences between gender for each education level. For example, if males attending university has a higher population mean job satisfaction score than females or another attending university. This can be seen on an interaction plot when traces over a segment are non-parallel.

4. Use lm to create a model for job satisfaction score using education level and gender as the predictors, including the interaction between education level and gender. Show the code to create the model. Create the usual 4 regression diagnostic plots. Include the plots in your report. Comment in about 60 words on whether there is anything that causes you concern about the least-squares linear model assumptions.

```
m1 <- lm(score ~ education_level * gender, data=my_js)
summary(m1)

##
## Call:
## lm(formula = score ~ education_level * gender, data = my_js)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12400 -0.49288  0.01185  0.43521  1.58000
##
## Coefficients:
##                                         Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                             6.36300    0.20518  31.012  < 2e-
16 ***
## education_levelschool                  -0.60800    0.30777  -1.976
0.0511 .
## education_leveluniversity               2.21100    0.29017   7.620 1.89e-
11 ***
## genderfemale                           -0.03838    0.27291  -0.141
0.8884
## gendermale                             -0.06383    0.27781  -0.230
0.8188
## education_levelschool:genderfemale     -0.37508    0.39936  -0.939
0.3500
## education_leveluniversity:genderfemale -0.54562    0.38944  -1.401
0.1645
## education_levelschool:gendermale       -0.11450    0.40606  -0.282
0.7786
## education_leveluniversity:gendermale    0.45698    0.38646   1.182
0.2400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6488 on 95 degrees of freedom
## Multiple R-squared:  0.8151, Adjusted R-squared:  0.7995
## F-statistic: 52.34 on 8 and 95 DF,  p-value: < 2.2e-16

par(mfrow = c(2,2))
plot(m1)
```



Linearity and no influential points assumptions are met.

In the Normal Q-Q plot, some residuals deviate from the line especially at the extremities. This is a concern as the residuals might not resemble a normal distribution.

In the Scale Location, the residuals are forming a zigzag pattern. This is a concern as this might not support equal variance within the residuals.

5. Show the analysis of variance for the model given by the R function anova. State the p-value for the F-test for the interaction between education level and gender shown on the `anova' output. Explain in about 80 words what probability this p-value represents.

```
anova(m1)

## Analysis of Variance Table
##
## Response: score
##                        Df  Sum Sq Mean Sq  F value  Pr(>F)
## education_level         2 169.172  84.586 200.9233 < 2e-16 ***
## gender                  2   3.420   1.710   4.0613 0.02030 *
## education_level:gender  4   3.689   0.922   2.1907 0.07583 .
## Residuals              95  39.994   0.421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the F-test for the interaction between education level and gender is 0.07583. The interaction term is encapsulating as it is the highest order term and its p-value would determine if we'll need to look at the lower order terms instead. This determines if the interaction has an effect on the model/shows variation. This p-value tests if the model without the interaction term explains the job satisfaction score as well as the model with the interaction term (null hypothesis).

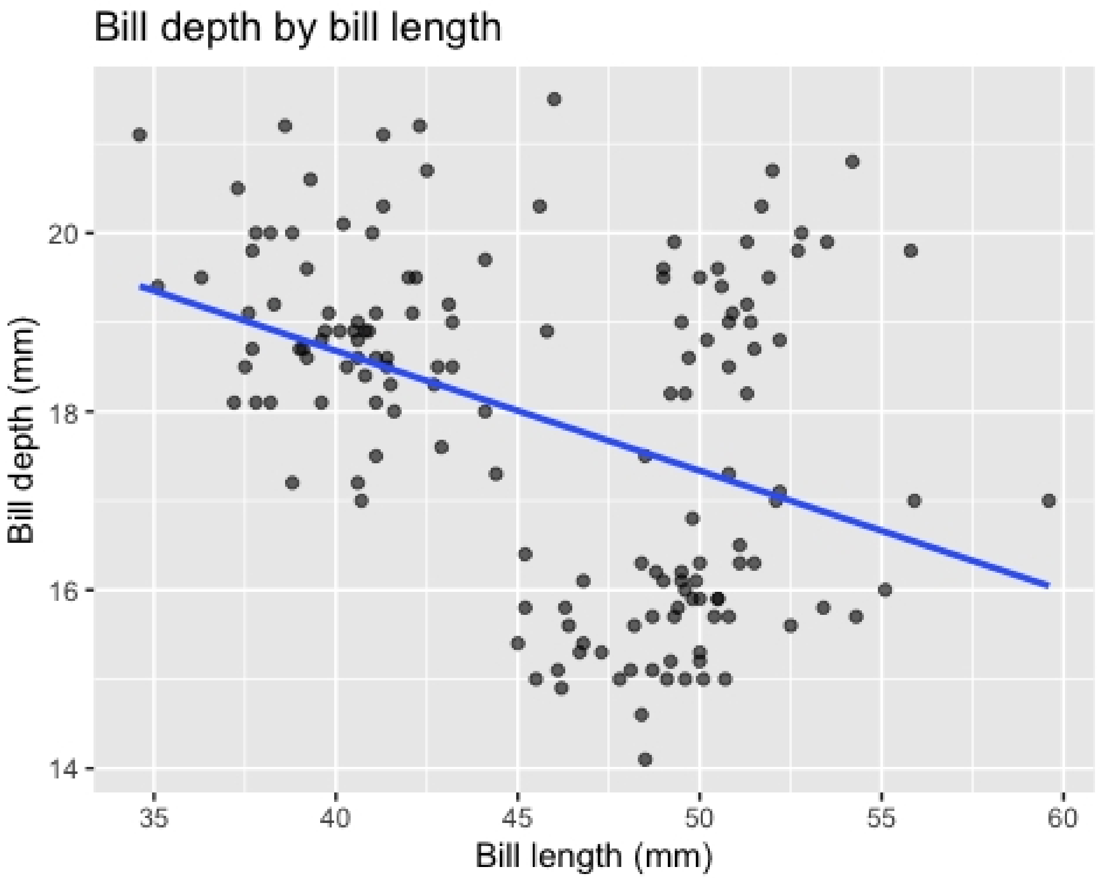
The p-value for the interaction term of 0.07583 is not statistically significant. There is enough evidence to reject the null hypothesis of non-interaction model explaining job score as well as the interaction model. Therefore drop the interaction and go for the model without the interaction.

6. Load the palmerpenguins library. You will also need tidyverse if it is not already loaded. Filter to get the male sample only. The following code will do both steps for you and create a sample called male_penguins of the male penguins with no "n/a" values.

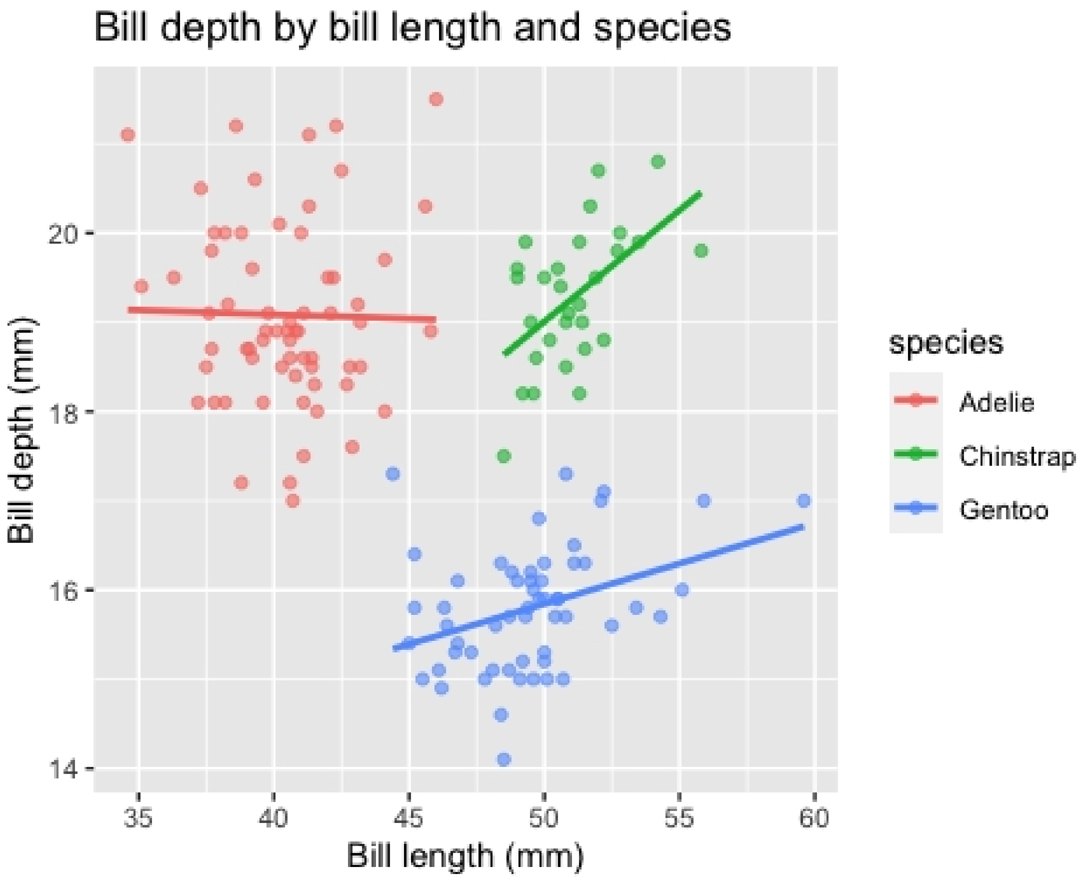
```
library(palmerpenguins)
male_penguins <- penguins %>%
  drop_na() %>%
  filter(sex == "male")

set.seed(31878039)
my_pen_m <- male_penguins %>%
  sample_n(150)
```

7. Use ggplot to create a scatter plot of bill_depth_mm (y-axis) against bill_length_mm (x-axis) but ignoring species (ignoring species is not a good idea, as we will see very shortly!). Use geom_smooth to plot the single linear regression line for response bill_depth_mm, predictor bill_length_mm. The following code will create the required scatter plot and regression line:

```
my_pen_m %>% ggplot(aes(
 x = bill_length_mm,
 y = bill_depth_mm
)) +
 geom_point(alpha = 0.6) +
 geom_smooth(method = lm, se = FALSE) +
   labs(
 title = "Bill depth by bill length",
 x = "Bill length (mm)",
 y = "Bill depth (mm)"
 )

## `geom_smooth()` using formula 'y ~ x'
```



```
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) +
   labs(
 title = "Bill depth by bill length and species",
 x = "Bill length (mm)",
 y = "Bill depth (mm)"
 )

## `geom_smooth()` using formula 'y ~ x'
```



The first plot shows a negative relationship implying that as bill length increases, bill depth decreases.

In contrast, the second plot shows three relationships Adelie specie shows an almost constant slope implying no change in the bill depth. Gentoo and Chinstrap species show a positive relationship implying that as bill length increases, bill depth increases as well. However, Chinstrap specie has a faster rate than Gentoo indicating that Chinstrap has greater bill depth than Gentoo.

8. You are going to create models for bill depth (response) using bill length and species as predictors with and without an interaction between bill length and species. Describe in about 40 words how the slope(s) of the regression line(s) using the model with interactions could differ from the slope(s) of the regression line(s) using the model without interactions. Create a model to predict male penguin bill depth using both bill length and species, including the interaction between bill length and species. Show the code to create the model in your report. Also create a model to predict male penguin bill depth using both bill length and species, but with no interaction between bill length and species. Show the code to create the model in your report.

```
m2 <- lm(bill_depth_mm ~ bill_length_mm * species, data=my_pen_m)
m3 <- lm(bill_depth_mm ~ bill_length_mm + species, data=my_pen_m)
```

Each specie in the non-interaction model will have the same slope coefficient, but different intercepts. In contrast, each specie in the interaction model will have its own slopes and own intercepts. Therefore, parallel slopes in interaction model whereas non-parallel slopes in non-interaction model.

9. Perform a nested model ANOVA F-test using R's anova function to compare the two models you created in part [8]. State in 60-80 words what null and alternative hypotheses this F-test is testing. Give your conclusion for this test and explain how the nested model ANOVA output justifies this conclusion (up to 50 words in total).

```
anova(m3, m2)

## Analysis of Variance Table
##
## Model 1: bill_depth_mm ~ bill_length_mm + species
## Model 2: bill_depth_mm ~ bill_length_mm * species
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    146 107.27
## 2    144 102.35  2    4.9262 3.4655 0.03389 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis of this F-test is that in the population model, the coefficients of all regressors in the full model are zero. This means that the slopes for each specie in the interaction model are zero. The alternative hypothesis is that in the population model, the coefficients of all regressors in the full model are not zero. This means that the slopes for each specie in the interaction model are not zero.

This F-test gives out a p-value of 0.03389 which is statistically significant. Therefore, reject the null hypothesis as there is enough evidence that at least one of the extra regressors in the interaction model is useful in explaining the bill depth (response).

10. Give the equations of the regression lines for each of the 3 species using the model including the interaction between bill length and species from part [8] (you will find it helpful to get R's summary output for the model to be able to work out the values of the intercepts and slopes for each species). Two decimal places for each coefficient is sufficient!

```
summary(m2)

## 
## Call:
## lm(formula = bill_depth_mm ~ bill_length_mm * species, data = my_pen_m)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.08041 -0.56844 -0.08042  0.46295  2.46983 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    19.46623    1.80769  10.769  < 2e-16 ***
## bill_length_mm                 -0.00948    0.04465  -0.212  0.83216    
## speciesChinstrap              -12.93966    5.19252  -2.492  0.01384 *  
## speciesGentoo                  -8.14122    2.69753  -3.018  0.00301 ** 
## bill_length_mm:speciesChinstrap  0.25910    0.10520   2.463  0.01496 *  
## bill_length_mm:speciesGentoo     0.09985    0.06023   1.658  0.09952 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8431 on 144 degrees of freedom
## Multiple R-squared:  0.7973, Adjusted R-squared:  0.7903 
## F-statistic: 113.3 on 5 and 144 DF,  p-value: < 2.2e-16
```

Adelie specie:

bill depth = 19.47 - 0.01(bill_length_mm)

Chinstrap specie: bill depth = (19.47 - 12.94) + (-0.01 + 0.26)(bill_length_mm)

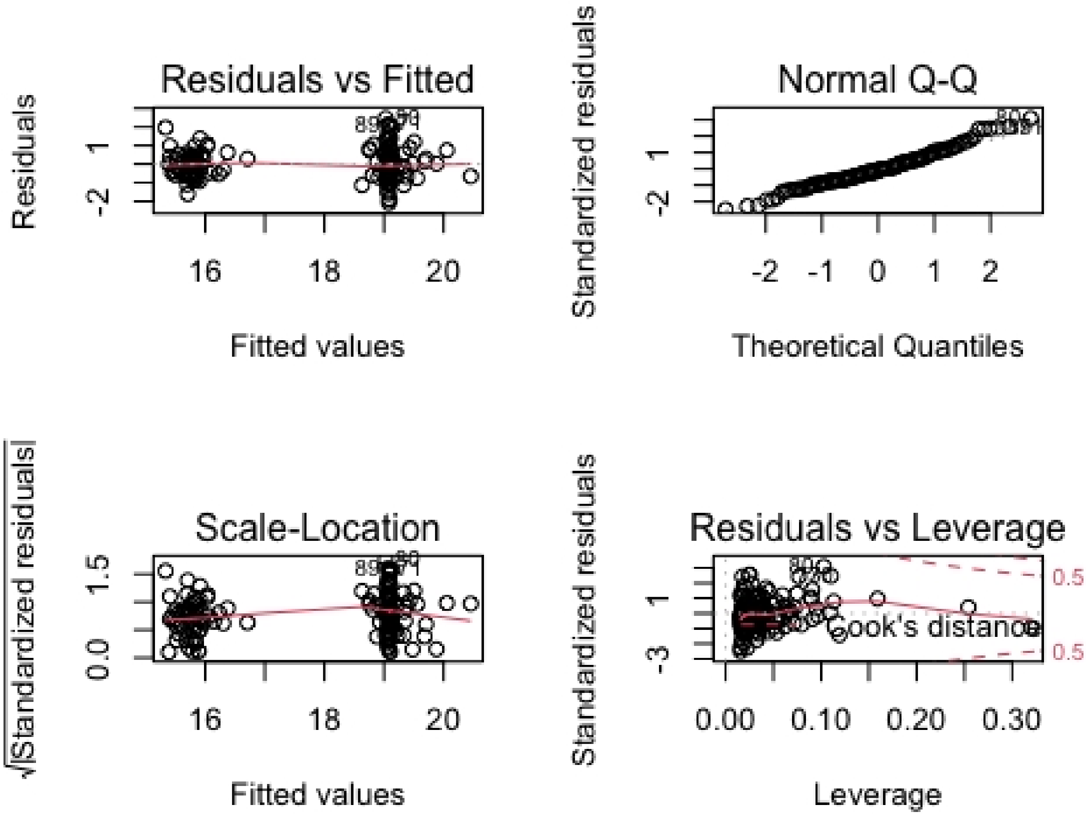
bill depth = 6.53 + 0.25(temp)

Gentoo specie: bill depth = (19.47 - 8.14) + (-0.01 + 0.10)(bill_length_mm)

bill depth = 11.33 + 0.09(temp)

11. Create the usual 4 regression diagnostic plots for the model including the interaction between bill length and species from part [8]. Include the plots in your report. Comment in 40-60 words on anything that causes you concern about the least-squares linear model assumptions.

```
par(mfrow = c(2,2))
plot(m2)
```



Residuals are linear as seen in Residuals vs Fitted.

Residuals follow the line with deviations at the extremities as seen in Normal Q-Q plot. This is a concern as the residuals could not be normally distributed.

Residuals have constant variance as they are scattered and have no pattern as seen in Scale-Location.

There’s no influential points as seen in Residuals vs Leverage.