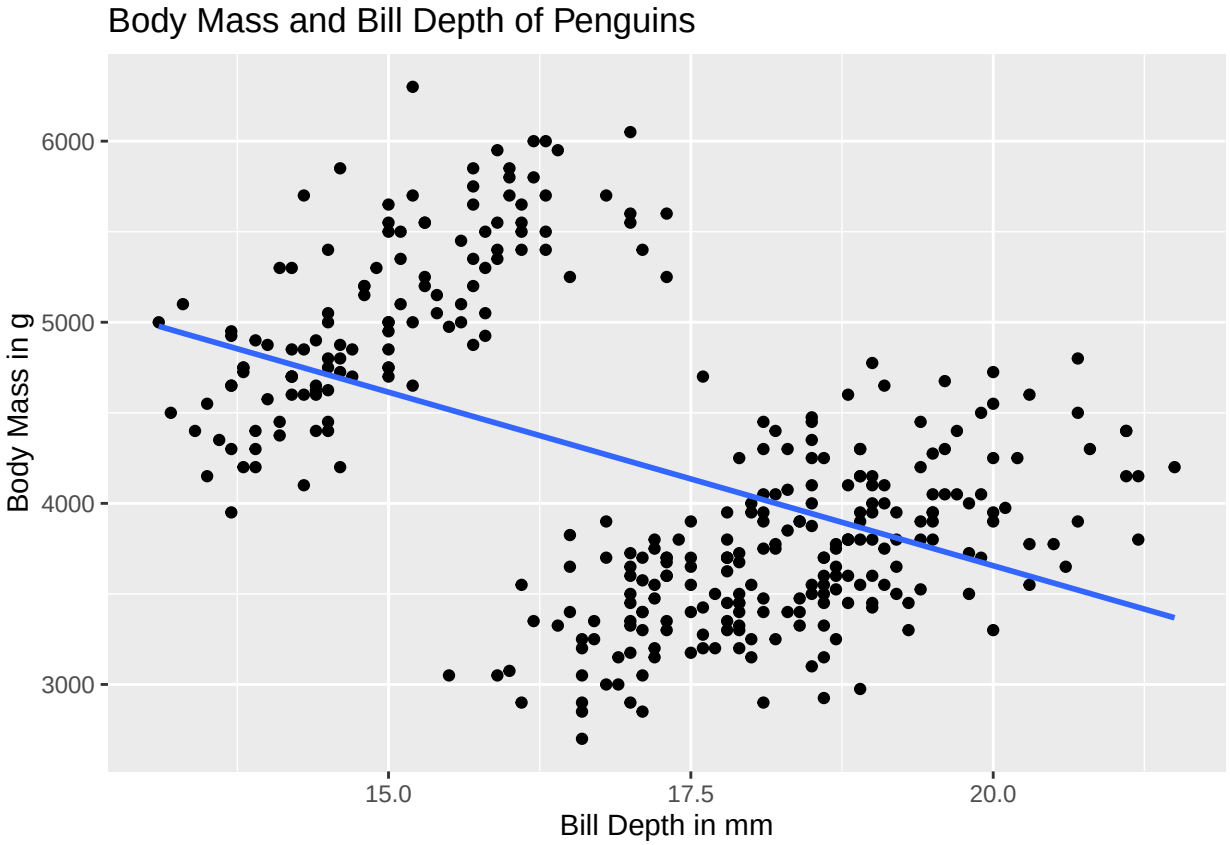
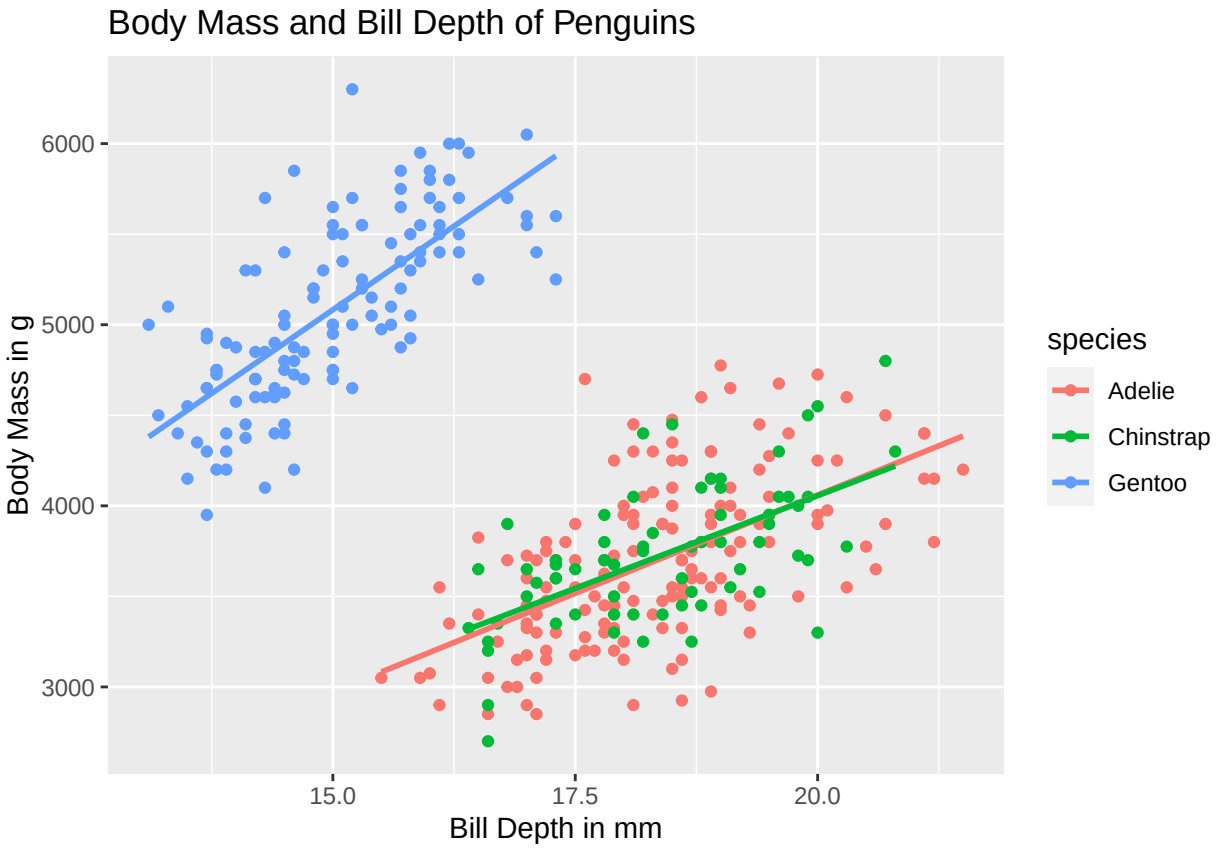
Guided Question Set 8 Solutions

1)



Based on this scatterplot, we note a negative linear association between bill depth and body mass of the penguins. Generally, as bill depth increases, body mass decreases. However, we do notice there appears to be two clusters of observations, which could reflect that the data are grouped in some manner.

2)



When we group the observations by species, we can see that within each species, there is a positive linear association between bill depth and body mass of the penguins. Generally, as bill depth increases, body mass increases.

We note that the relationship between bill depth and body mass appears to be almost identical for Adelie and Chinstrap penguins, since their regression lines are parallel. The slope of the regression line for Gentoo penguins is a bit steeper, indicating a slightly larger increase in body mass for each unit change in bill depth, on average.

3)

```
contrasts(Data$species)
```

```
##           Chinstrap Gentoo
## Adelie            0      0
## Chinstrap         1      0
## Gentoo            0      1
```

We note that we are asked to make Adelie penguins the reference class, which we confirm via the `contrasts()` function.

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm * species, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -911.18 -251.93  -31.77  197.82 1144.81
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -2535.837    879.468  -2.883  0.00419 **
## flipper_length_mm                    32.832      4.627   7.095 7.69e-12 ***
## speciesChinstrap                   -501.359   1523.459  -0.329  0.74229
## speciesGentoo                     -4251.444   1427.332  -2.979  0.00311 **
## flipper_length_mm:speciesChinstrap     1.742      7.856   0.222  0.82467
## flipper_length_mm:speciesGentoo       21.791      6.941   3.139  0.00184 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 370.6 on 336 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7896, Adjusted R-squared:  0.7864
## F-statistic: 252.2 on 5 and 336 DF,  p-value: < 2.2e-16
```

The estimated regression equation is

$$\hat{y} = -2535.837 + 32.832x_1 - 501.359I_1 - 4251.444I_2 + 1.742x_1I_1 + 21.791x_1I_2,$$

where $I_1 = 1$ for Chinstrap penguins and 0 otherwise, and $I_2 = 1$ for Gentoo penguins and 0 otherwise.

Do note that the estimated coefficients for β_2 and β_4 are insignificant, which indicates that the intercepts and slopes for Adelie and Chinstrap penguins are not significantly different.

4)

```
## Analysis of Variance Table
##
## Model 1: body_mass_g ~ flipper_length_mm + species
## Model 2: body_mass_g ~ flipper_length_mm * species
##   Res.Df      RSS Df Sum of Sq     F   Pr(>F)
## 1    338 47666988
```

```
## 2    336 46147424  2   1519564 5.532 0.004327 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \beta_4 = \beta_5 = 0$, H_a : at least one of the coefficients in H_0 is not zero.

Since the p-value of this partial F test is small, we reject the null hypothesis. We cannot drop both interaction terms.

The relationship between bill depth and body mass is not the same for all three species (not surprising given the different slope for Gentoo penguins).

5)

From part 3, we know that the estimated regression equation is

$$\hat{y} = -2535.837 + 32.832x_1 - 501.359I_1 - 4251.444I_2 + 1.742x_1I_1 + 21.791x_1I_2,$$

where $I_1 = 1$ for Chinstrap penguins and 0 otherwise, and $I_2 = 1$ for Gentoo penguins and 0 otherwise. So plugging in the values for the indicators, we have

For Adelie penguins,

$$\begin{aligned}\hat{y} &= -2535.837 + 32.832x_1 - 501.359 \times 0 - 4251.444 \times 0 + 1.742x_1 \times 0 + 21.791x_1 \times 0 \\ &= -2535.837 + 32.832x_1.\end{aligned}$$

For Chinstrap penguins,

$$\begin{aligned}\hat{y} &= -2535.837 + 32.832x_1 - 501.359 \times 1 - 4251.444 \times 0 + 1.742x_1 \times 1 + 21.791x_1 \times 0 \\ &= -2535.837 - 501.359 + (32.832 + 1.742)x_1 \\ &= -3037.196 + 34.574x_1.\end{aligned}$$

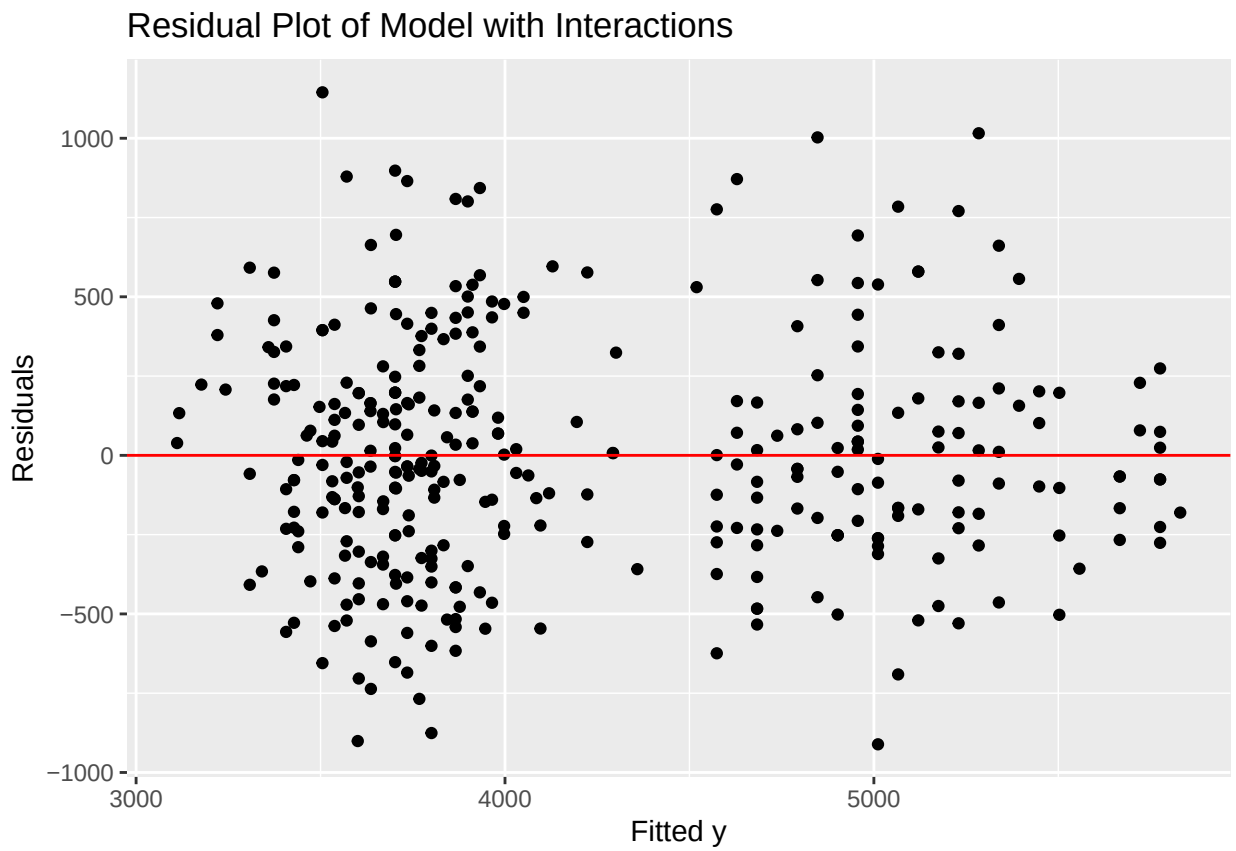
For Gentoo penguins,

$$\begin{aligned}\hat{y} &= -2535.837 + 32.832x_1 - 501.359 \times 0 - 4251.444 \times 1 + 1.742x_1 \times 0 + 21.791x_1 \times 1 \\ &= -2535.837 - 4251.444 + (32.832 + 21.791)x_1 \\ &= -6787.281 + 54.623x_1.\end{aligned}$$

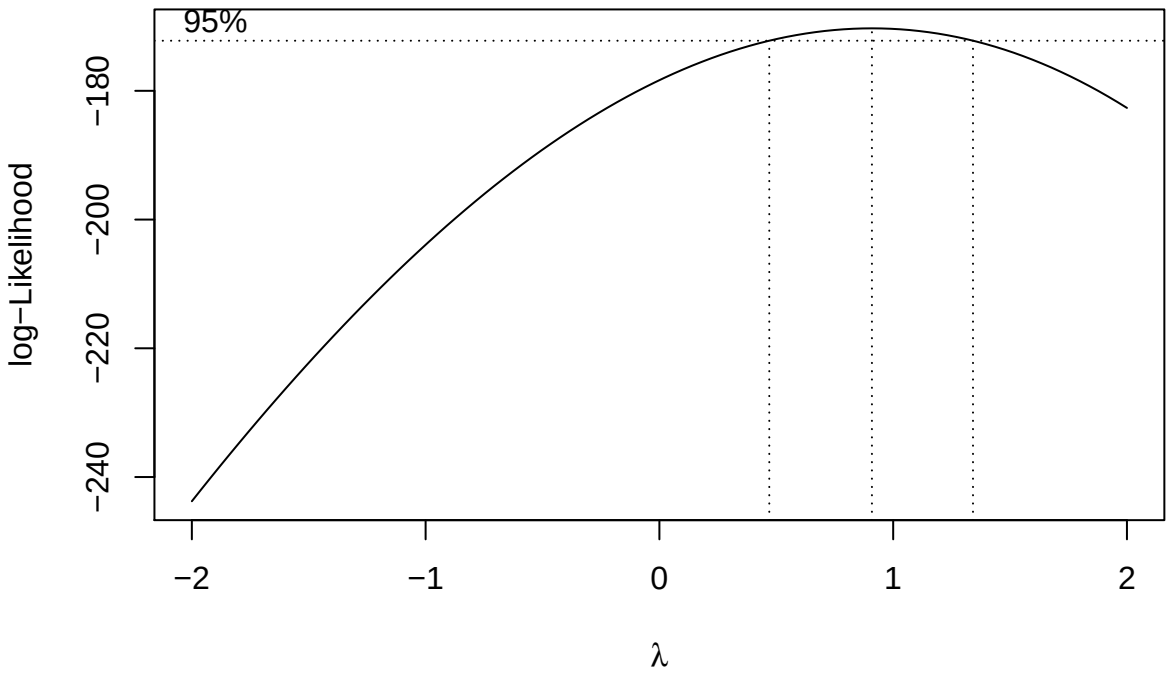
As mentioned earlier, the estimated coefficients for β_2 and β_4 are insignificant, which indicates that the intercepts and slopes for Adelie and Chinstrap penguins are not significantly different.

6)

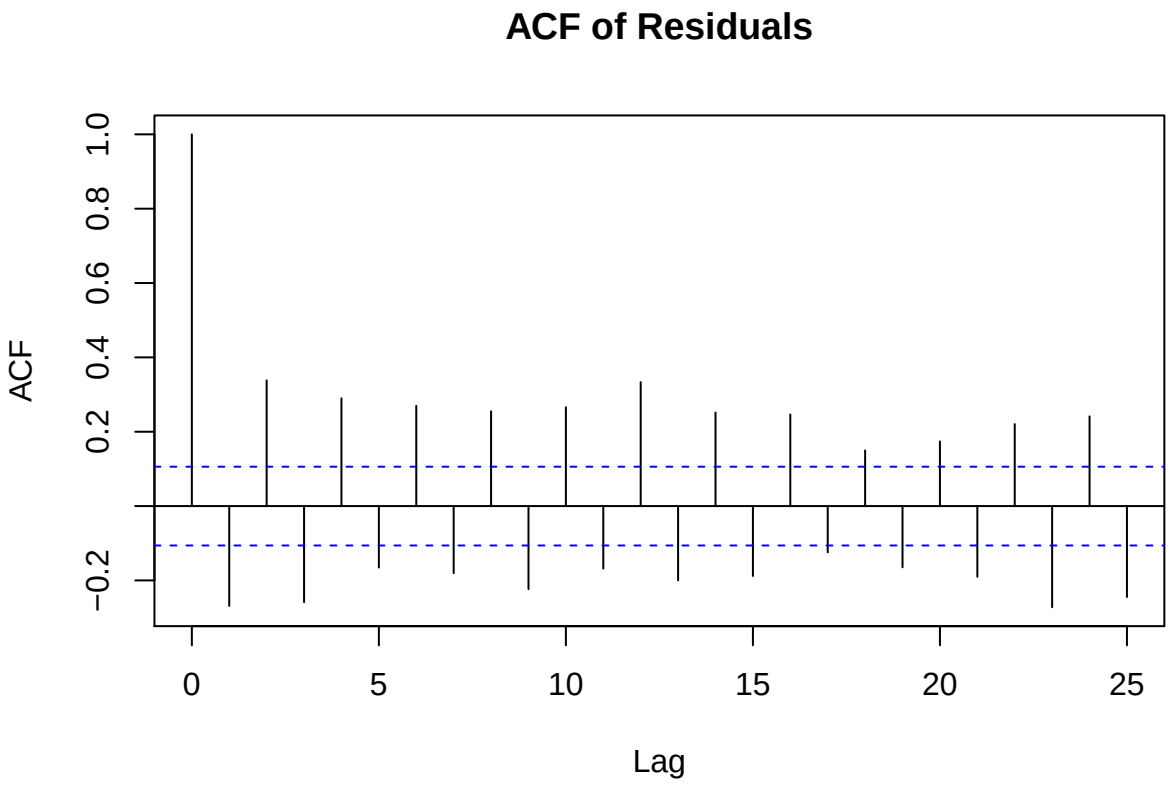
First we create the residual plot. We see that the residuals are evenly scattered acorss the horizontal axis, so the mean of the residuals is 0. We also note the vertical spread of the residuals is fairly constant, so the constant variance assumption is met.



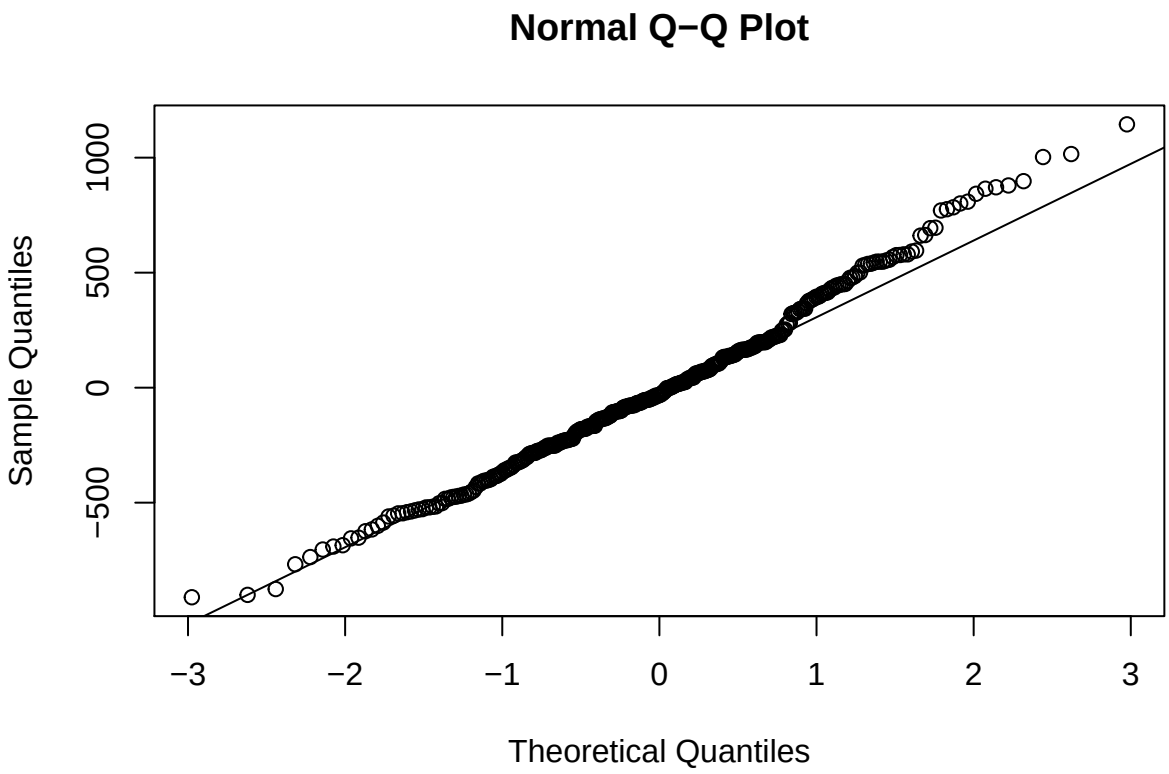
Next, we take a look at the Box Cox plot to confirm our observation that the variance is constant and hence we do not need to transform the respones variable. Since the value of 1 lies in the CI, we do not need to transform the response variable.



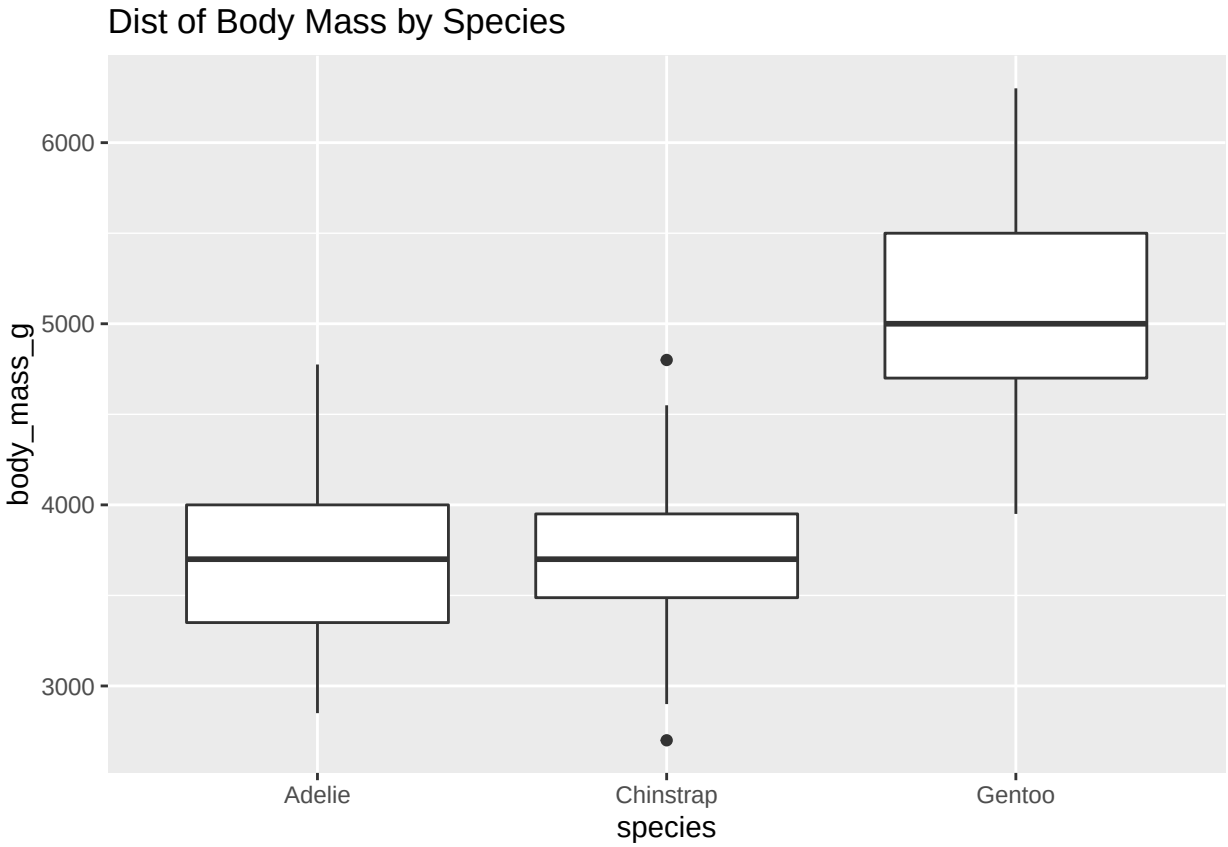
Next, we look at the ACF plot of the residuals. We notice that the residuals are significantly correlated. Looking back at the original data set, we notice that the data are sorted by species, and then island. We have establised the body masses differ by species, so it is not surprising that our residuals are correlated.



Next, we look at the QQ plot of the residuals. The normality assumption is met since the plots fall close to the straight line.



Finally, we check if the variance of the response variable is constant across all species. The boxplots indicate the spread of body mass is similar across all species, and Levene's test for equality of variances across all species was insignificant, so we don't have evidence the variances are not the same across all species.




```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  Data$body_mass_g
## Test Statistic = 2.8123, p-value = 0.09468
```

7)

Since our partial F test informs us that we should use the model with interactions between species and bill length, the relationship between bill length and body mass differs across species. This was noted by the different slopes in the scatterplot in part 2.

To carry out Tukey's multiple comparisons, we need the slopes to be equal (or no significant interactions), as multiple comparisons assume the differences in body mass between species is the same as long as bill depth is controlled for.