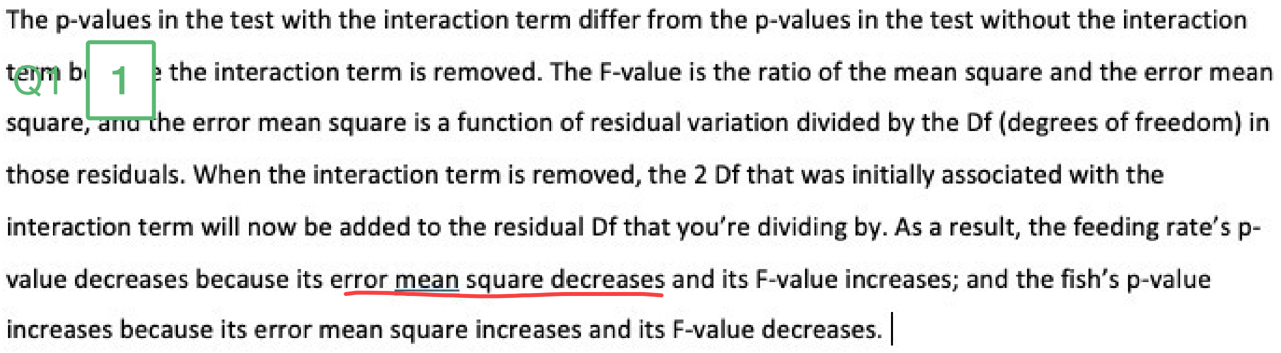
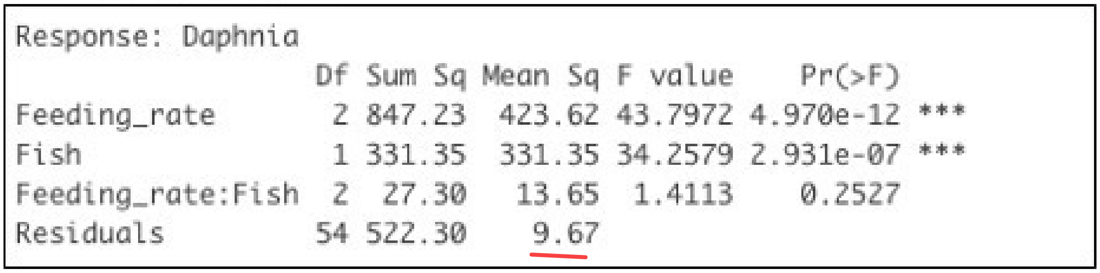


Q1 1

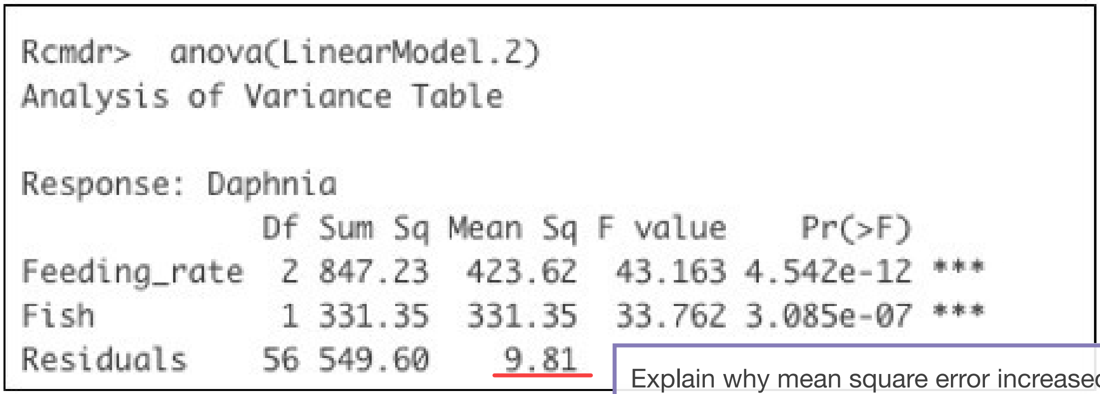
The p-values in the test with the interaction term differ from the p-values in the test without the interaction term because the interaction term is removed. The F-value is the ratio of the mean square and the error mean square, and the error mean square is a function of residual variation divided by the Df (degrees of freedom) in those residuals. When the interaction term is removed, the 2 Df that was initially associated with the interaction term will now be added to the residual Df that you're dividing by. As a result, the feeding rate's p-value decreases because its error mean square decreases and its F-value increases; and the fish's p-value increases because its error mean square increases and its F-value decreases.

With interaction term:



```
Response: Daphnia
                  Df Sum Sq Mean Sq F value    Pr(>F)
Feeding_rate       2 847.23  423.62 43.7972 4.970e-12 ***
Fish               1 331.35  331.35 34.2579 2.931e-07 ***
Feeding_rate:Fish  2  27.30   13.65  1.4113    0.2527
Residuals         54 522.30    9.67
```

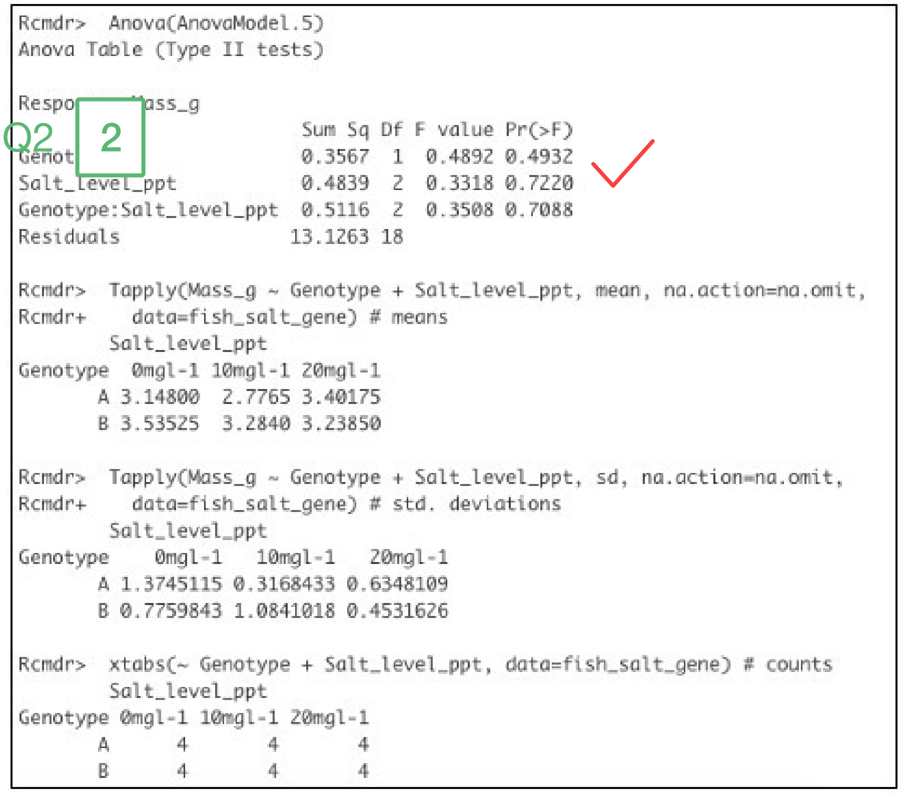
Without interaction term:



```
Rcmdr>  anova(LinearModel.2)
Analysis of Variance Table

Response: Daphnia
             Df Sum Sq Mean Sq F value    Pr(>F)
Feeding_rate  2 847.23  423.62  43.163 4.542e-12 ***
Fish          1 331.35  331.35  33.762 3.085e-07 ***
Residuals    56 549.60    9.81
```

Explain why mean square error increased, despite larger df typically causing a smaller MSE

Rcmdr output after doing multi-way ANOVA for fish_salt_gene:



Q2 2

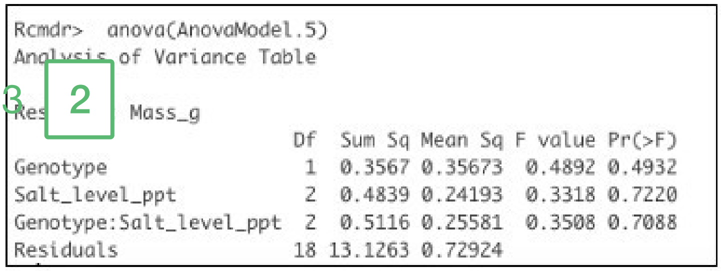
```
Rcmdr>  Anova(AnovaModel.5)
Anova Table (Type II tests)

Response: Mass_g
                         Sum Sq Df F value Pr(>F)
Genotype                 0.3567  1  0.4892 0.4932
Salt_level_ppt           0.4839  2  0.3318 0.7220
Genotype:Salt_level_ppt  0.5116  2  0.3508 0.7088
Residuals               13.1263 18

Rcmdr>  Tapply(Mass_g ~ Genotype + Salt_level_ppt, mean, na.action=na.omit,
Rcmdr+    data=fish_salt_gene) # means
        Salt_level_ppt
Genotype  0mgl-1 10mgl-1 20mgl-1
       A 3.14800  2.7765 3.40175
       B 3.53525  3.2840 3.23850

Rcmdr>  Tapply(Mass_g ~ Genotype + Salt_level_ppt, sd, na.action=na.omit,
Rcmdr+    data=fish_salt_gene) # std. deviations
        Salt_level_ppt
Genotype    0mgl-1   10mgl-1   20mgl-1
       A 1.3745115 0.3168433 0.6348109
       B 0.7759843 1.0841018 0.4531626

Rcmdr>  xtabs(~ Genotype + Salt_level_ppt, data=fish_salt_gene) # counts
        Salt_level_ppt
Genotype 0mgl-1 10mgl-1 20mgl-1
       A      4       4       4
       B      4       4       4
```

✓

It shows that the p-values of the genotype and salt level are both greater than 0.05, with genotype p-value=0.4932 and salt level p-value=0.7220. This implies that neither factors have a significant effect on the fish mass and therefore, running another linear model is not necessary.

✓

Q3 2



```
Rcmdr>  anova(AnovaModel.5)
Analysis of Variance Table

Res        Mass_g
                        Df  Sum Sq Mean Sq F value Pr(>F)
Genotype                 1  0.3567 0.35673  0.4892 0.4932
Salt_level_ppt           2  0.4839 0.24193  0.3318 0.7220
Genotype:Salt_level_ppt  2  0.5116 0.25581  0.3508 0.7088
Residuals               18 13.1263 0.72924
```

The p-value for interaction term of genotype and salt level is 0.7088. This p-value is greater than 0.05, hence, the interaction between the salinity and genotype is not significant. Since the individual factors (i.e., genotype and salinity) and interaction terms are all insignificant, re-running the linear model is not necessary. ✓

Q4 2

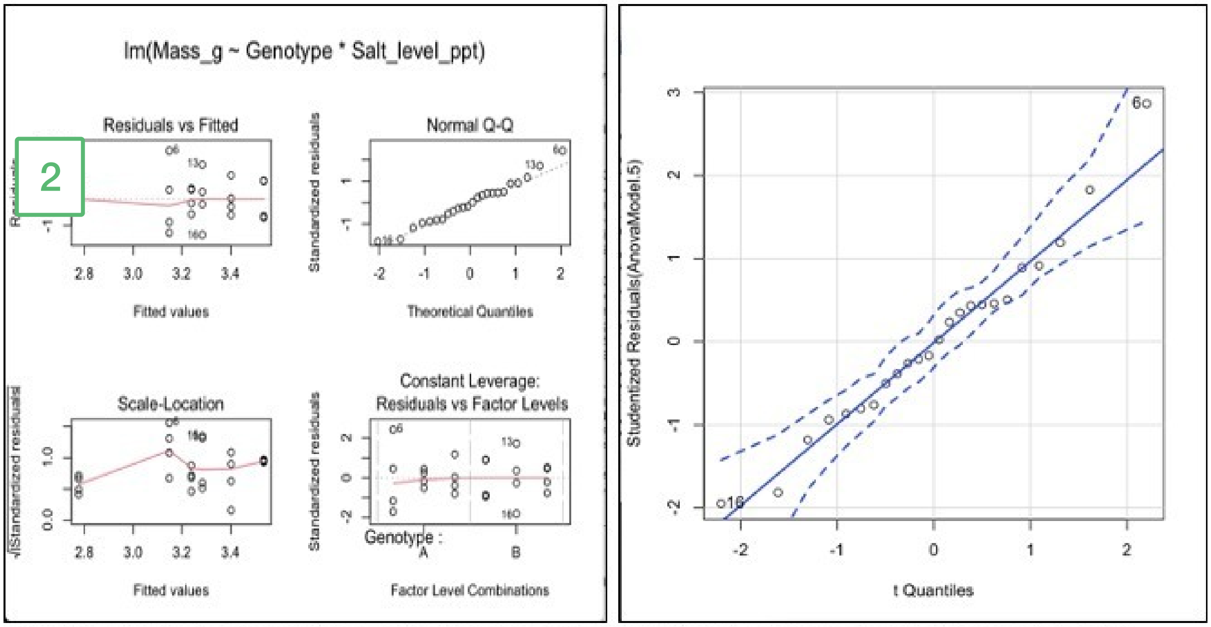


Figure 1. The basic diagnostic plots to check for the distribution of residuals among the genotype and salinity (left). The residual quantile comparison plot of genotype and salinity (right). Both graphs show normality of the residuals. ✓

The residuals follow normality because the red line in the Residuals vs Fitted graph falls within the normal line. In addition, the normal QQ plot shows the linearity – most of the data points fall within the normal line. Moreover, on the graph on the right, we can see that the data points show a pattern of linearity and lie within the 95% compatibility interval. For these reasons, we can conclude that the residuals are normally distributed.

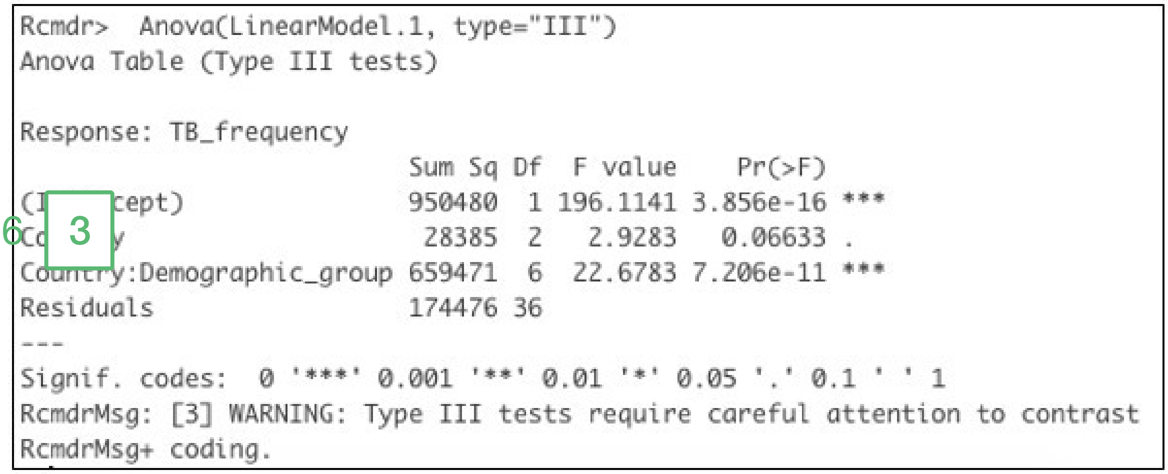
Q5 1.75

5. Multi-way ANOVA was used to determine whether salinity, genotype, and the interaction of both salinity and genotype affect the mass of yearling fish. Using type I multiway ANOVA, it was determined that the p-valu[illegible]type, salinity and the interaction term between the genotype and salinity are 0.4932, 0.72[illegible]088, respectively. In addition, it was also determined that the degrees of freedom (Df) for genotype, salt level, interaction term and residuals are 1, 2, 2 and 18, respectively.

   Since all the p-values are greater than the alpha value of 0.05, it implies that neither the genotype, salinity nor the interaction between the genotype and salinity have an effect on the yearling fish mass. Moreover, it also implies that there is no necessity to re-run a linear model as none of these factors have been determined to have no significant effect on the mass of the fish. Doing so will yield the same p-value results.

   Thus, it can be concluded that the mass of the yearling fish is not affected regardless of its genotype, the salinity of water or the combination of both salinity and genotype.

Q6 3

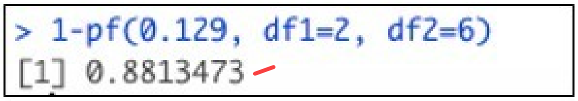


```
Rcmdr>  Anova(LinearModel.1, type="III")
Anova Table (Type III tests)

Response: TB_frequency
                           Sum Sq Df  F value    Pr(>F)    
(Intercept)                950480  1 196.1141 3.856e-16 ***
Country                     28385  2   2.9283   0.06633 .  
Country:Demographic_group  659471  6  22.6783 7.206e-11 ***
Residuals                  174476 36                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RcmdrMsg: [3] WARNING: Type III tests require careful attention to contrast
RcmdrMsg+ coding.
```

Using a scientific calculator, the corrected F-value of the country is calculated to be:

F-value =(Country Sum of Squares/Country Degrees of freedom)*(Interaction term Degrees of freedom/Interaction term Sum of Squares) = (28385/2)*(6/659471) = 0.12912622 = 0.129

Using the following formula on the R console, the test significance, i.e., p-value, of this F-value is calculated to be:

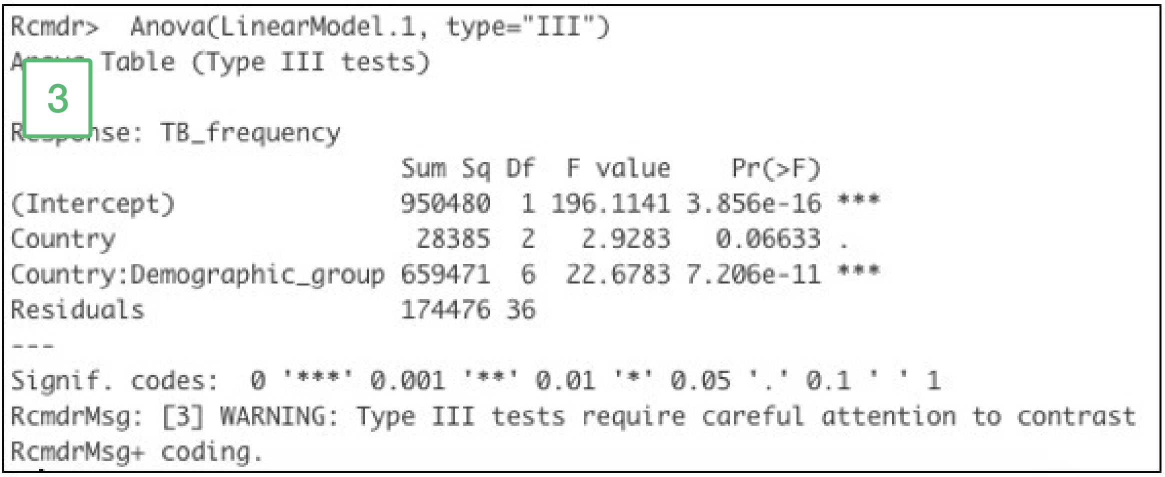


```
> 1-pf(0.129, df1=2, df2=6)
[1] 0.8813473
```

The corrected F-value for the Country is 0.129 and the corrected p-value for the Country is 0.881. Since the corrected Country F-value is notably small and the corrected Country p-value is greater than 0.05, there is sufficient evidence that there is no difference in the frequency of tuberculosis among the different countries.

2.

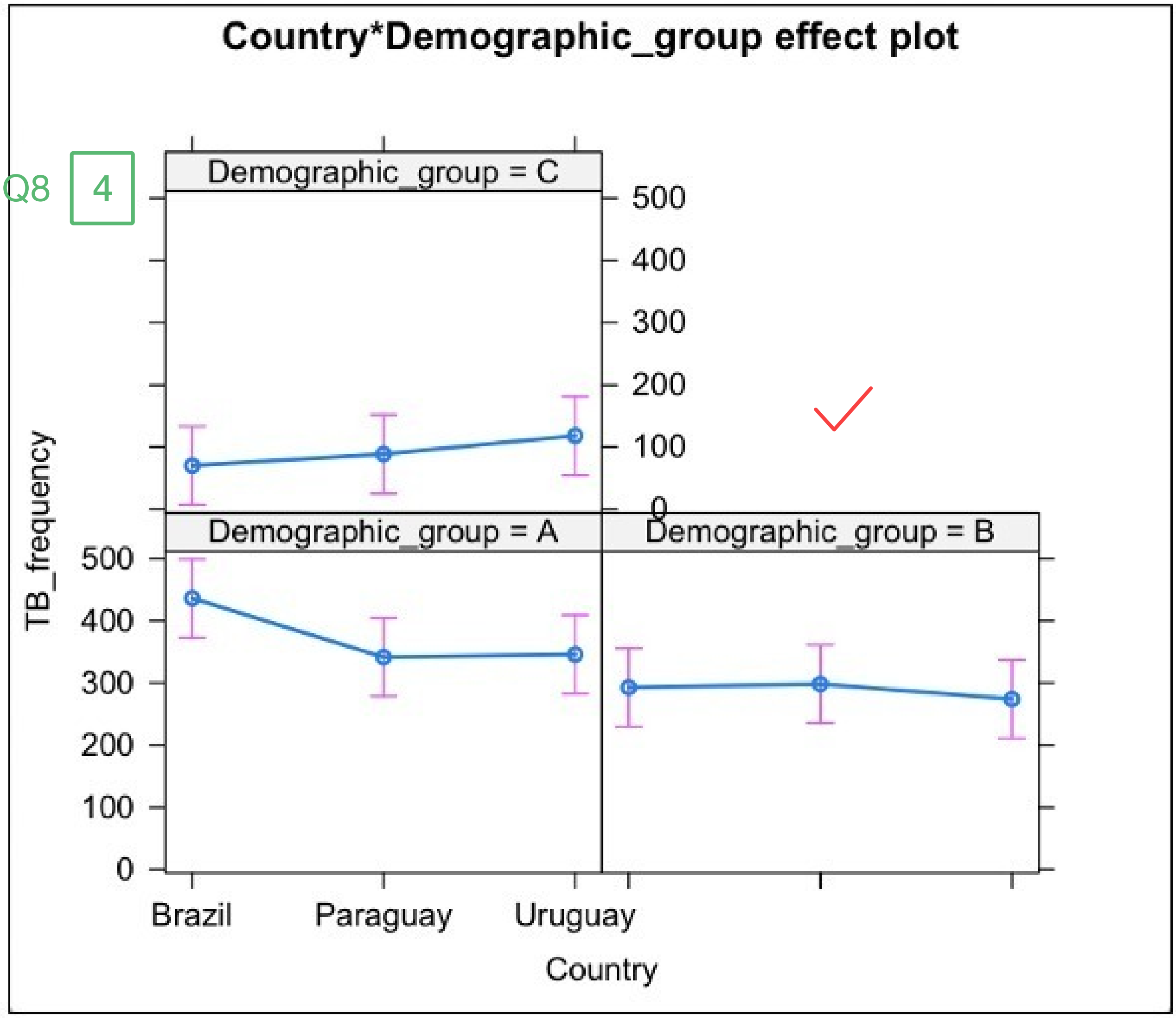
Q7 3



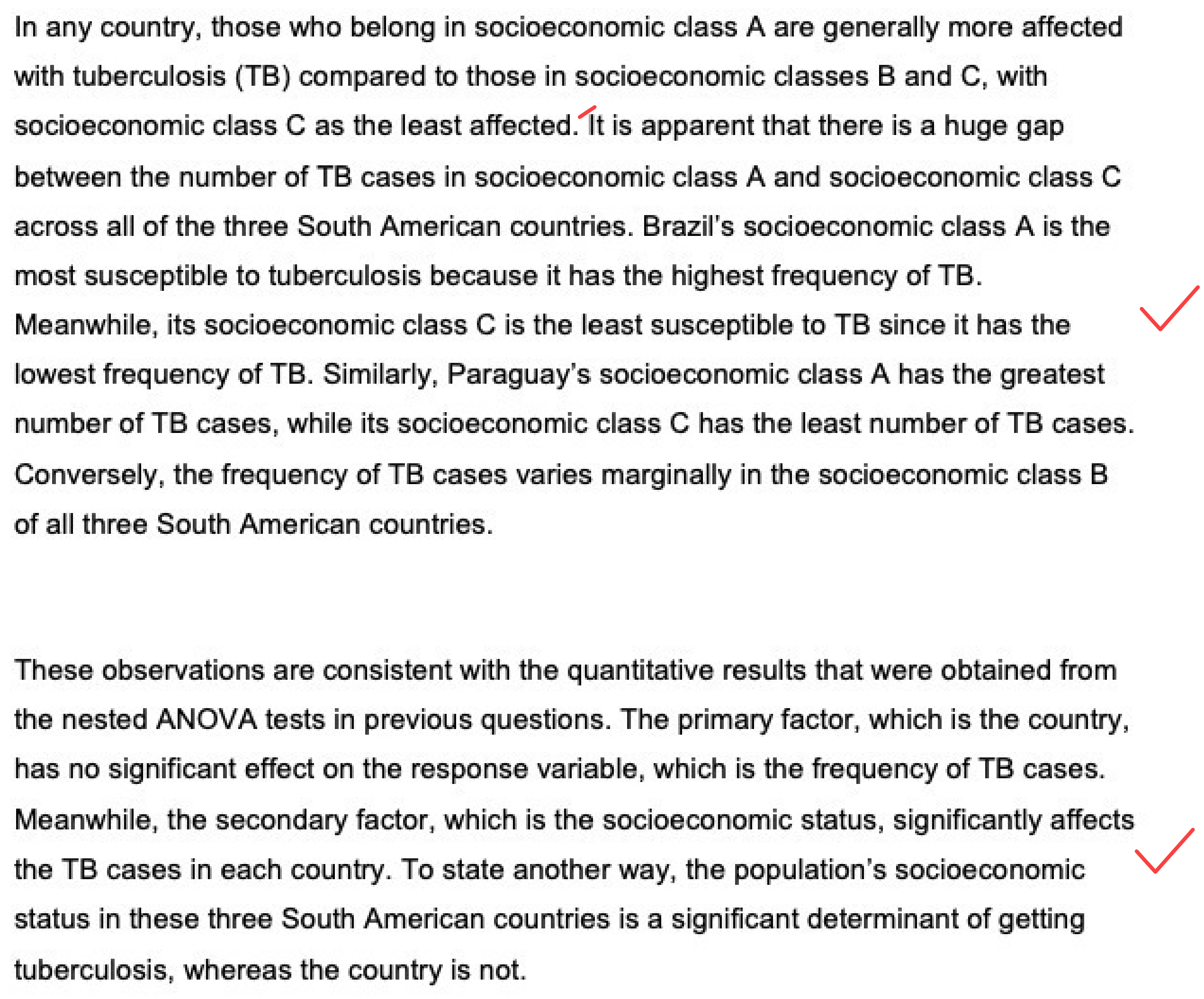
```
Rcmdr>  Anova(LinearModel.1, type="III")
Anova Table (Type III tests)

Response: TB_frequency
                          Sum Sq Df  F value    Pr(>F)    
(Intercept)               950480  1 196.1141 3.856e-16 ***
Country                    28385  2   2.9283   0.06633 .  
Country:Demographic_group 659471  6  22.6783 7.206e-11 ***
Residuals                 174476 36                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RcmdrMsg: [3] WARNING: Type III tests require careful attention to contrast
RcmdrMsg+ coding.
```

The F-value of the socioeconomic groups within different countries is identified to be 22.6783, with p-value 7.206e-11. The F-value is considerably a large number and the p-value is less than 0.05, hence there is difference in the tuberculosis frequency among socioeconomic groups within the different countries.

Q8 4


Country*Demographic_group effect plot
Demographic_group = C
500
400
300
200
100
0
Demographic_group = A
Demographic_group = B
TB_frequency
500
400
300
200
100
0
Brazil
Paraguay
Uruguay
Country

✓



In any country, those who belong in socioeconomic class A are generally more affected with tuberculosis (TB) compared to those in socioeconomic classes B and C, with socioeconomic class C as the least affected. It is apparent that there is a huge gap between the number of TB cases in socioeconomic class A and socioeconomic class C across all of the three South American countries. Brazil's socioeconomic class A is the most susceptible to tuberculosis because it has the highest frequency of TB. Meanwhile, its socioeconomic class C is the least susceptible to TB since it has the lowest frequency of TB. Similarly, Paraguay's socioeconomic class A has the greatest number of TB cases, while its socioeconomic class C has the least number of TB cases. Conversely, the frequency of TB cases varies marginally in the socioeconomic class B of all three South American countries.

These observations are consistent with the quantitative results that were obtained from the nested ANOVA tests in previous questions. The primary factor, which is the country, has no significant effect on the response variable, which is the frequency of TB cases. Meanwhile, the secondary factor, which is the socioeconomic status, significantly affects the TB cases in each country. To state another way, the population's socioeconomic status in these three South American countries is a significant determinant of getting tuberculosis, whereas the country is not.