

Assignment2 - STAT8121

Minh Tien Ta - 46207031

20 Oct 2021

Contents

Question1	1
Question2	9

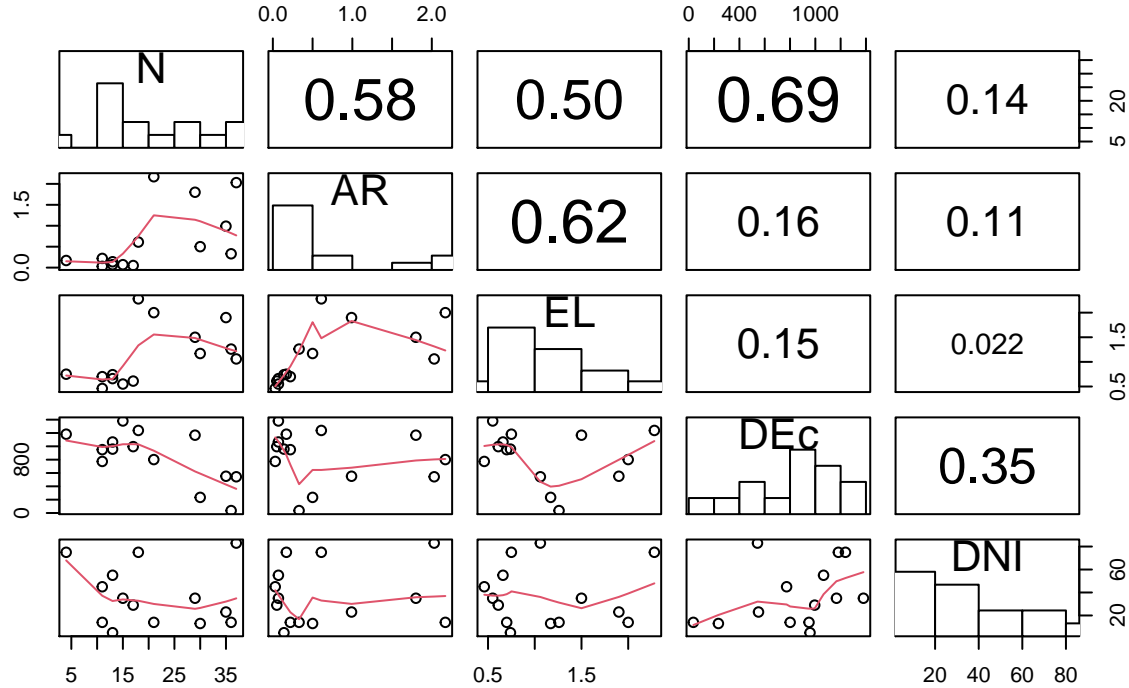
Question1

Reading dataset

a. Produce a scatterplot and correlation matrix of the data and comment on possible relationships between the response and predictors and relationships between the predictors themselves.

```
pairs(paramo[1:5],  
      upper.panel = panel.cor, # show correlation ratio  
      diag.panel  = panel.hist, # show histogram  
      lower.panel = panel.smooth, # show the line  
      main = "the correlation of between the variables"  
      # pch = "."  
      )
```

the correlation of between the variables



Comments:

- There are moderate correlations between N and AR, EL, DEc (0.58, 0.5, 0.69) respectively.
- Meanwhile, the correlation of other variables is slightly positive.

b. Conduct an F-test for the overall regression

b1. Write down the mathematical multiple regression model for this situation, defining all appropriate parameters

- We have the multiple regression model which can be written by this:

Regression line: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon; \quad \varepsilon \sim N(0, \sigma^2)$

where:

- Y : N - The number of species of birds observed.
- X_1, X_2, X_3, X_4 : AR, EL, DEc, DNI variables respectively.
- $\beta_1, \beta_2, \beta_3, \beta_4$: AR, EL, DEc, DNI regression coefficients respectively.
- β_0 : Intercept term

b2. Write down the Hypotheses for the Overall ANOVA test of multiple regression

- Hypotheses of Anova test in multiple regression:

$H_0 : \beta_1, \beta_2, \beta_3, \beta_4 = 0; \text{ and } H_1 = \text{at least one } \beta_i \neq 0$

b3. Produce an ANOVA table for the overall multiple regression model (One combined regression SS source is sufficient)

Now, Anova tables would be:

```
paramo.aov = anova(lm(N ~ AR + EL + DEc + DNI , data = paramo))
paramo.aov

## Analysis of Variance Table
##
## Response: N
##          Df Sum Sq Mean Sq F value    Pr(>F)
## AR          1  508.92   508.92  11.3208 0.008328 **
## EL          1   45.90    45.90   1.0211 0.338661
## DEc          1  537.39   537.39  11.9541 0.007189 **
## DNI          1    2.06     2.06   0.0457 0.835412
## Residuals    9  404.59    44.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b4. Compute the F statistic for this test

$F_{statistic} = 6.085434$

```
n = nrow(paramo) ; k = ncol(paramo)
df1 = k - 1; df2 = n - k
Full_RegSS = sum(paramo.aov[["Mean Sq"]][1:4])
Reg_MS = Full_RegSS/4
Res_MS = paramo.aov[["Mean Sq"]][5]
F_obs = Reg_MS/ Res_MS ; F_obs
```

```
## [1] 6.085434
```

b5. State the Null distribution

- Null hypothesis suggests that the area of the island, elevation, distance from Ecuador, distance towards other islands does not affect the number of species of present birds.
- Alternative hypothesis suggests either area of the island or elevation or distance from Ecuador or distance to other islands or all predictors affect the number of species of birds present.
- If $p_{value} \leq \alpha$ reject the null hypothesis. $p_{value} > \alpha$ If fail to reject the null hypothesis

b6. Compute the P-Value

- P-value $P(F_{4,10} \geq 6.0854) = 0.0095 < 0.05$, then we reject H_0 at the 5% level.

```
pf(F_obs, df1, df2, lower.tail = FALSE)
```

```
## [1] 0.01182024
```

b7. State your conclusion (both statistical conclusion and contextual conclusion)

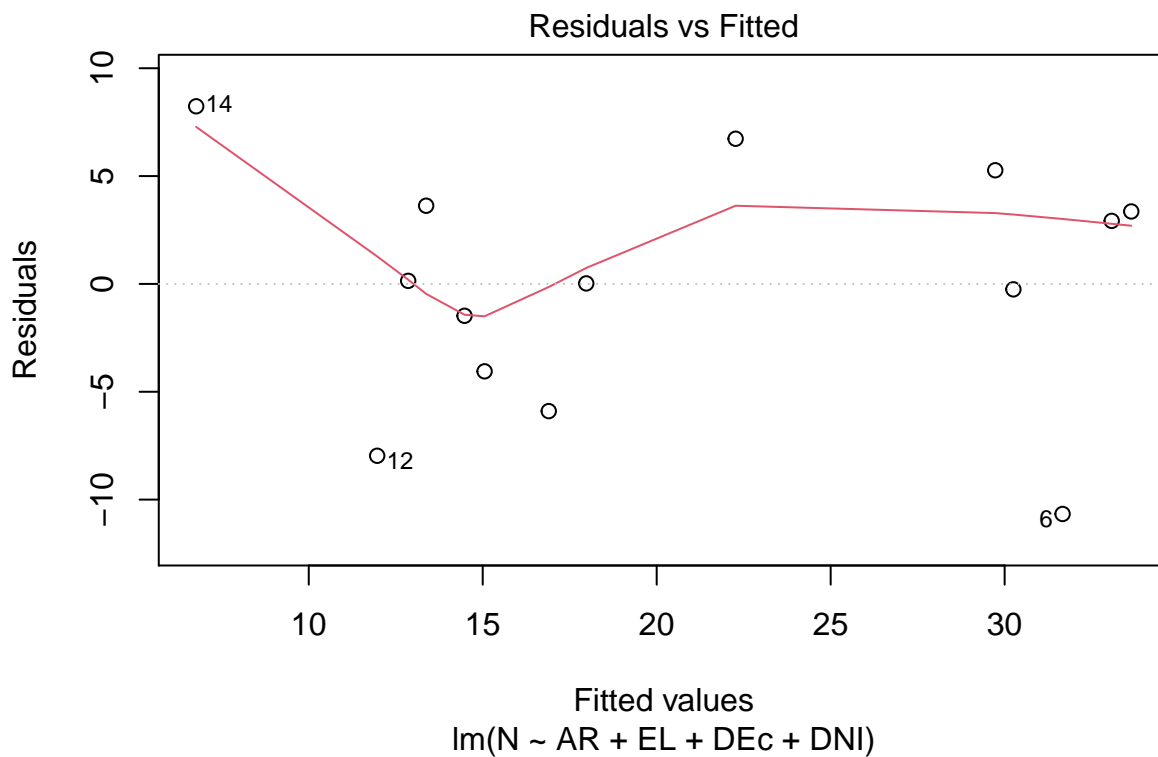
- There is a significant linear relationship between percentage response N and at least one of the four predictor variables.

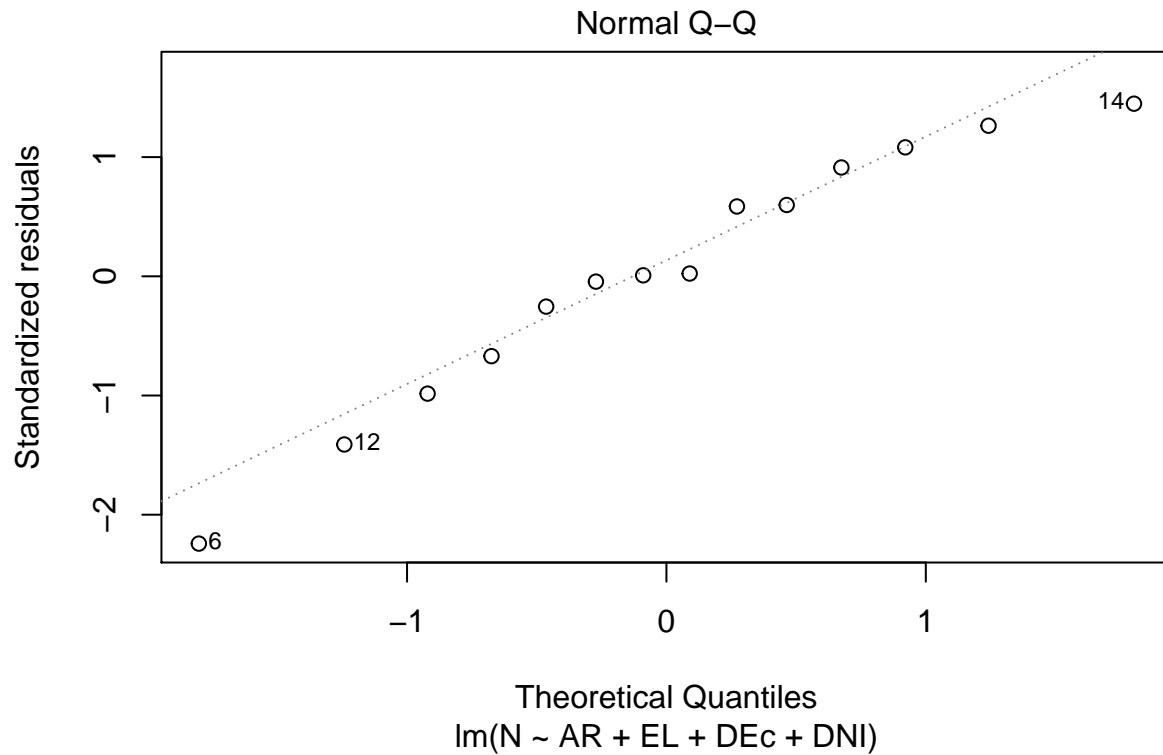
c. Validate the full model using all the predictors and comment on whether it is appropriate to a multiple regression model to explain the N abundance value.

- Now, we will check the assumptions of the model whether is appropriate to a multiple regression model

Check diagnostics:

```
plot(paramo.lm.all, which = 1:2)
```

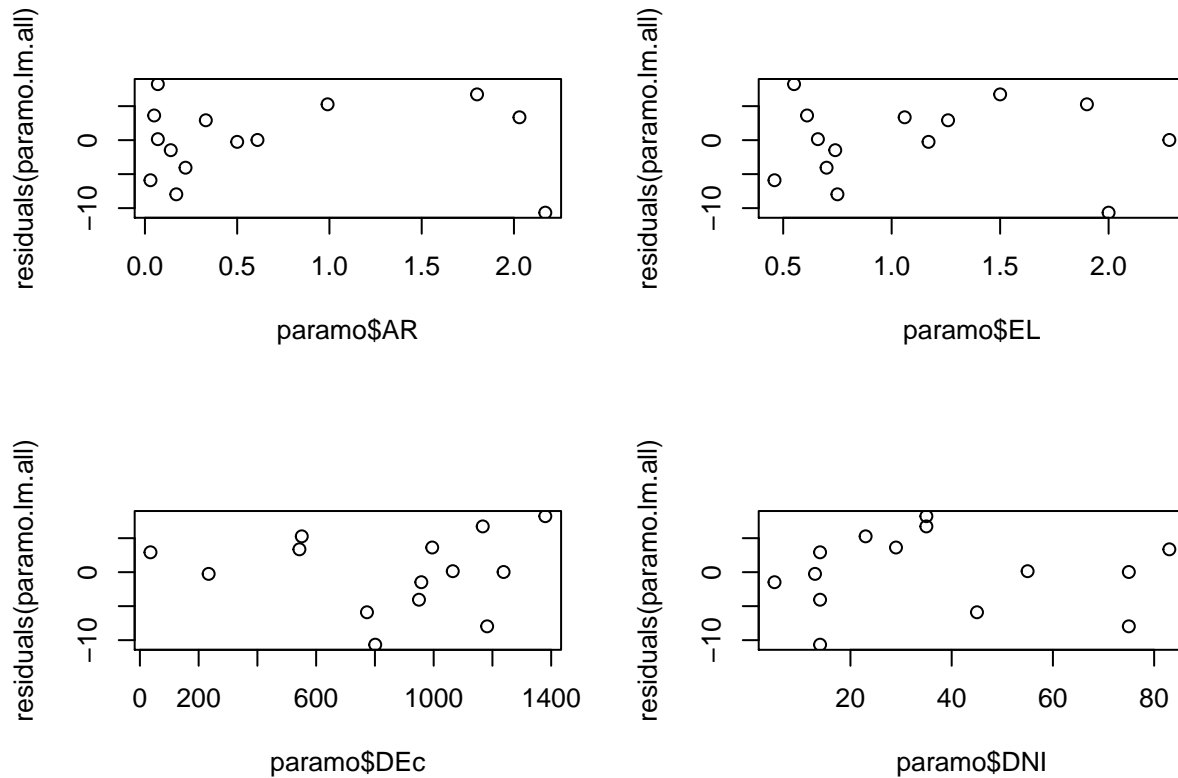




```
# Change line color and fill color
#create histogram of residuals
# ggplot(data = paramo, aes(x = paramo.lm.all$residuals)) +
#   geom_histogram(bins = 10,fill = 'steelblue', color = 'black') +
#   labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')
```

Check residuals against predictors:

```
# plot(resid(paramo.lm.all) ~ AR + EL + DEc + DNI, data= paramo)
par(mfrow = c(2, 2))
plot(paramo$AR, residuals(paramo.lm.all))
plot(paramo$EL, residuals(paramo.lm.all))
plot(paramo$DEc, residuals(paramo.lm.all))
plot(paramo$DNI, residuals(paramo.lm.all))
```



Comment:

1. The Normal Q-Q plot of residuals has slight bias on the top line but close to linear implying errors close to normally distributed.
2. The residuals vs fitted plot includes a spike in the smoothed plot, but the points has not discernable pattern.
3. Residuals vs predictor plots no obvious pattern. So linear model seems adequate.

d Find the R^2 and comment on what it means in the context of this dataset

We can obtain R^2 for the multiple regression like this:

- $R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{SS_{Total} - SS_{Residuals}}{SS_{Total}}$
- $R^2 = \frac{1498.857 - 404.5895}{1498.857} = 0.7301$

Comment:

- For linear regression models, R-squared is a measure of goodness-of-fit. .
- So, $R^2 = 0.7301$ which would indicate that 73% of the variance of the response variable being studied is explained by the variance of the predictor variable.

```
total_SumSQ = sum(paramo.aov[["Sum Sq"]][1:5])
SQ_reg = paramo.aov[["Sum Sq"]][5]
round((total_SumSQ - SQ_reg)/total_SumSQ,4)
```

```
## [1] 0.7301
```

e. Using model selection procedures used in the course, find the best multiple regression model that explains the data. State the final fitted regression model.

- Based on the result of the model in 4 predictors, we can remove the insignificant variable DNE

```
para.lm.2 = lm(formula = N ~ AR + EL + DEc , data = paramo)
summary(para.lm.2)
```

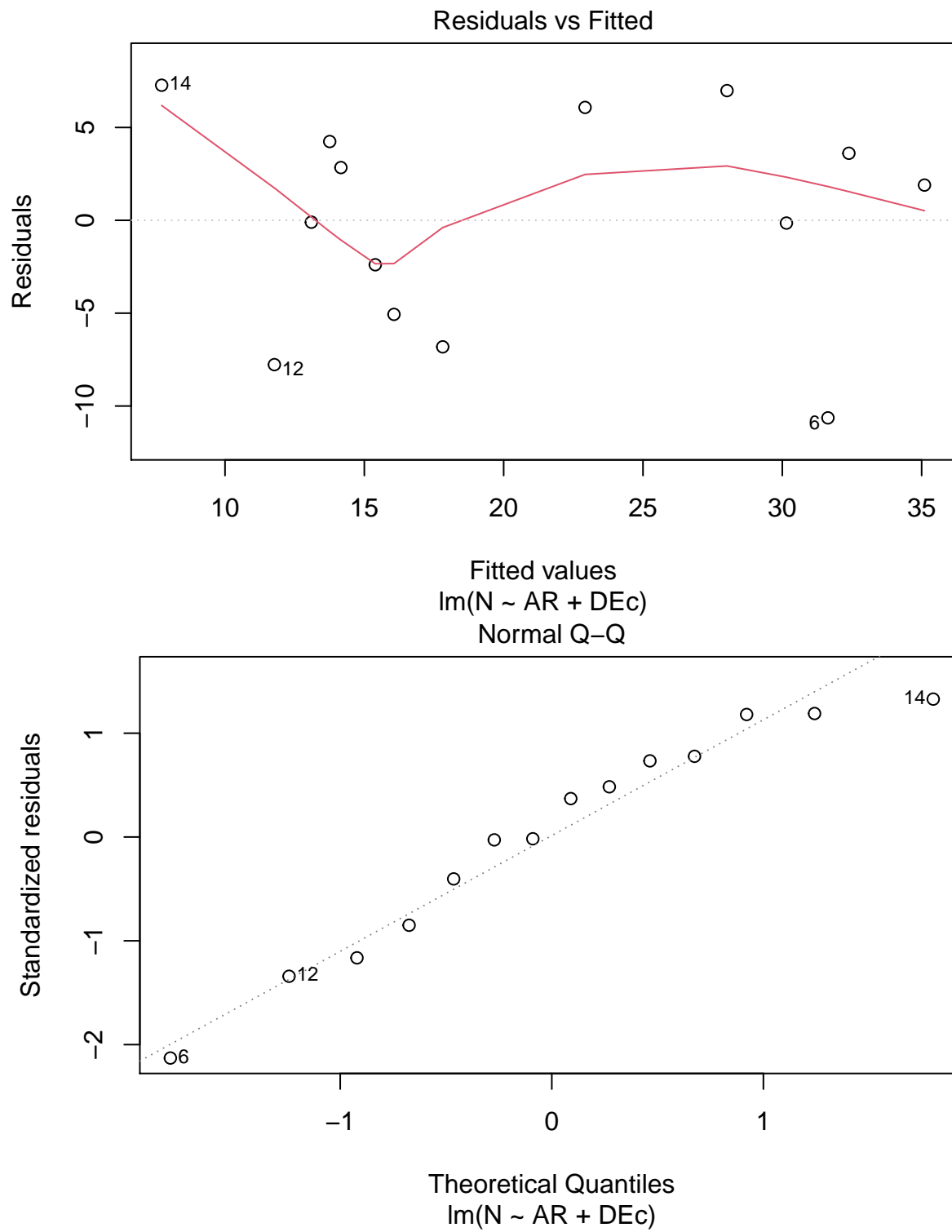
```
##
## Call:
## lm(formula = N ~ AR + EL + DEc, data = paramo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1638  -3.8306   0.4693   3.9477   8.0285
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.10415    5.80141   4.844 0.000677 ***
## AR           5.26428    2.90535   1.812 0.100087
## EL           3.04394    3.80214   0.801 0.441977
## DEc         -0.01679    0.00462  -3.635 0.004572 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.377 on 10 degrees of freedom
## Multiple R-squared:  0.7287, Adjusted R-squared:  0.6473
## F-statistic: 8.953 on 3 and 10 DF,  p-value: 0.003499
```

- Next, we determine that the EL variable is insignificant and eliminate it from the model.

```
para.lm.3 = lm(formula = N ~ AR + DEc , data = paramo)
summary(para.lm.3)
```

```
##
## Call:
## lm(formula = N ~ AR + DEc, data = paramo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6372  -4.3960   0.8989   4.0845   7.2734
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.797969   4.648155   6.626 3.73e-05 ***
## AR           6.683038   2.264403   2.951 0.01318 *
## DEc         -0.017057   0.004532  -3.764 0.00313 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.272 on 11 degrees of freedom
## Multiple R-squared:  0.7113, Adjusted R-squared:  0.6588
## F-statistic: 13.55 on 2 and 11 DF,  p-value: 0.001077
```

```
plot(para.lm.3, which = 1:2)
```



Comment:

- Therefore, the final model is fitted on the predictor variables as AR and DEc for explaining the response N.

f. Comment on the R^2 and Adjusted R^2 in the full and final model you chose in part e. In particular explain why those goodness of fitness measures change but not in the same way.

- Default adjusted R^2 for the model $lm(N \sim AR + DEc)$ is:
- Adjusted $R^2 = R^2 - (1 - R^2) \frac{p-1}{n-p} = 0.6588$ where $n = 14$ and $p = 3$ for model 3.

Comment:

- Adjusted R-squared is a variation of R-squared that has been modified to account for the model's predictor count. The adjusted R-squared value increases when the new term improves the model more than would be anticipated by chance and decreases when the new term improves the model less than would be expected by chance.
- So, we have the adjusted R^2 for the full model $lm(N \sim AR + EL + DEc + DNI)$ is 0.611 which is less than that of model $lm(N \sim AR + DEc)$. In other words, as compared to the full model, the model $lm(N \sim AR + DEc)$ improves the default R squared.

```
para.anova.3 = anova(lm(formula = N ~ AR + DEc , data = paramo))
R_2_model3 = (sum(para.anova.3$`Sum Sq`[1:3]) - para.anova.3$`Sum Sq`[3]) / sum(para.anova.3$`Sum Sq`[1:3])
p = 3; n = 14
AdjR_2_model3 = R_2_model3 - (1-R_2_model3)* (p-1)/(n-p) ; AdjR_2_model3
```

```
## [1] 0.6588177
```

g. Compute a 95% confidence interval for the AR regression parameter and explain what it means in the context of this data

- So, 95% confidence interval for AR is:
- $\widehat{\beta}_{AR} \pm t_{0.05,11} \times s.e(\widehat{\beta}_{AR}) = 6.683 \pm 2.032 \times 2.2644 = (1.6991, 11.6669)$

```
coef_AR = summary(para.lm.3)$coeff[2] ;
se_AR = summary(para.lm.3)$coef[5] ;
alpha = 0.05; t_table = qt(1- alpha/2,11) ;
c(coef_AR - t_table*se_AR,coef_AR + t_table*se_AR)
```

```
## [1] 1.699121 11.666955
```

Comment:

- We are 95% confident that the slope β_{AR} of the population regression line is between 1.699121 and 11.666955. We conclude that that of each additional of AR unit, then the N response will increase by between 1.69 and 11.66.

Question2

a. For this study, is the design balanced or unbalanced? Explain why

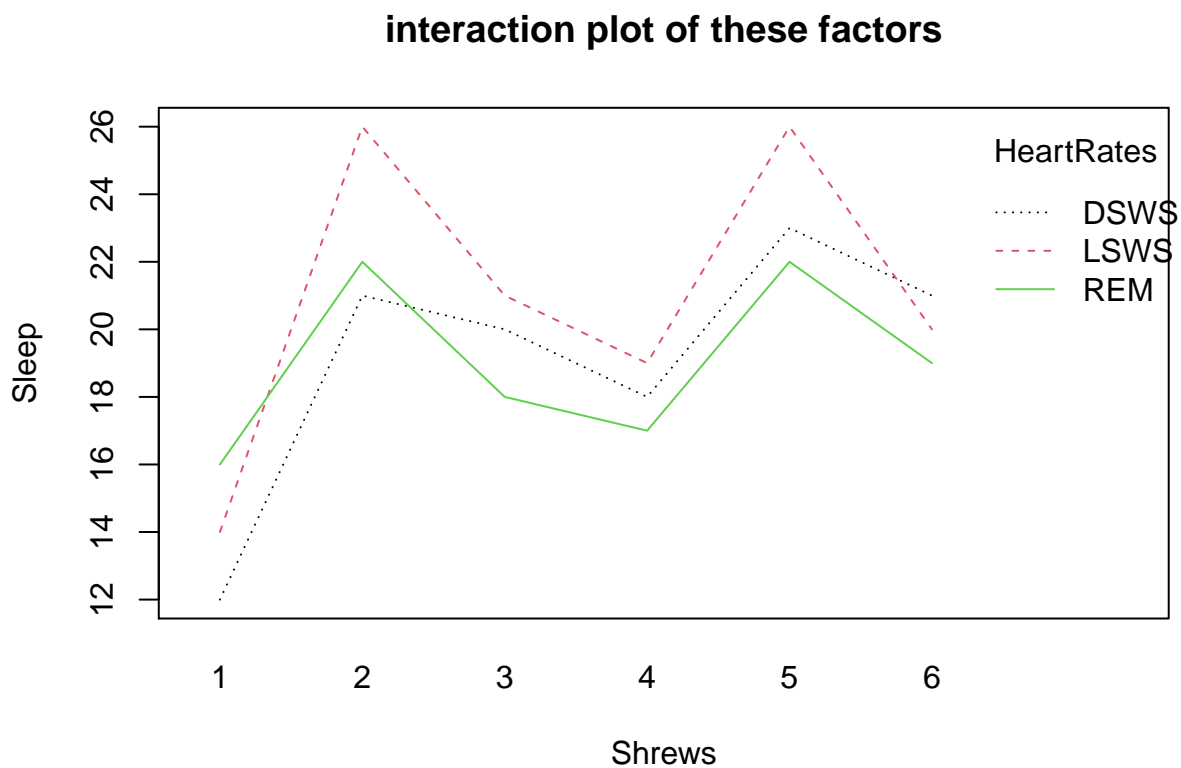
```
table(trShrew[, c("Shrews", "Sleep")])
```

```
##      Sleep
## Shrews DSWS LSWS REM
##      1   1   1   1
##      2   1   1   1
##      3   1   1   1
##      4   1   1   1
##      5   1   1   1
##      6   1   1   1
```

- As we can see the result, the design is balanced because of equal group sizes.

b. Construct two different preliminary graphs that investigate different features of the data and comment.

```
# Preliminary Investigation
with(trShrew, interaction.plot(Shrews, Sleep, HeartRates,
trace.label = "HeartRates",xlab = "Shrews", ylab = "Sleep",main = "interaction plot of these factors",c
```

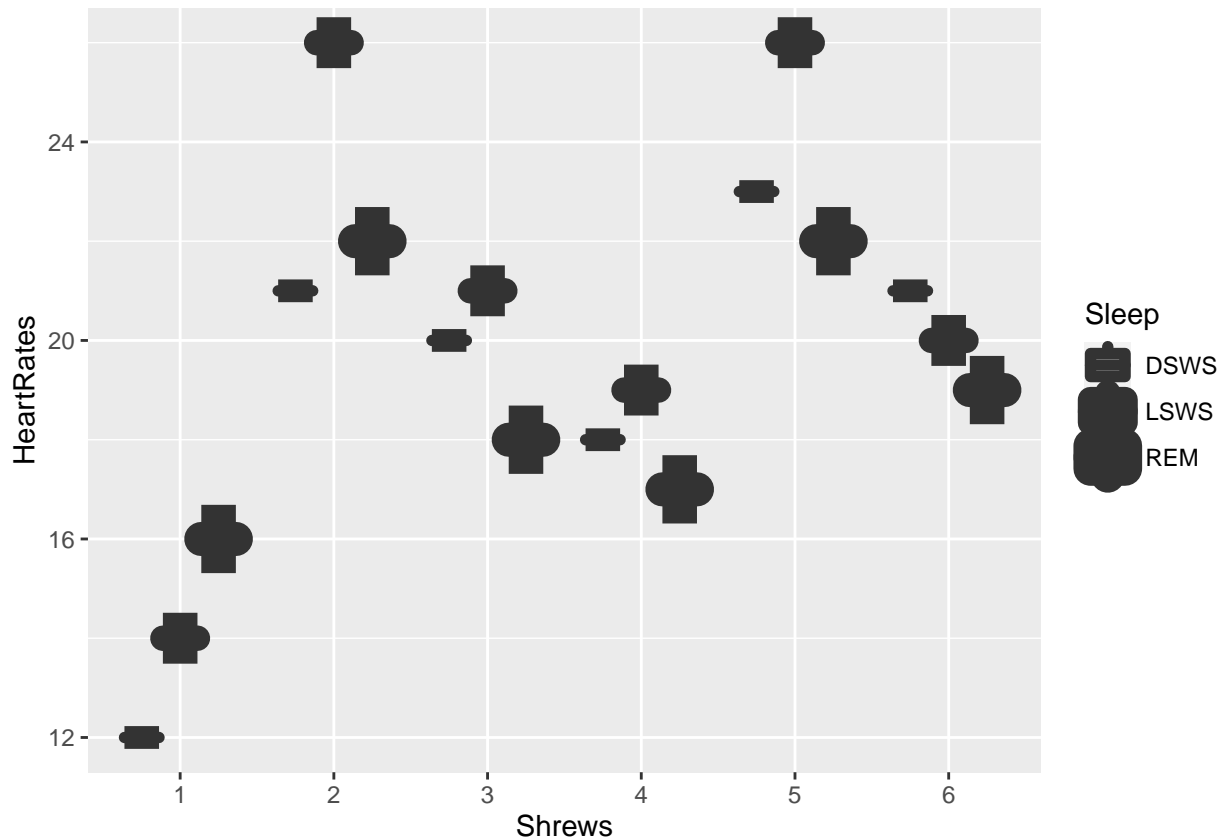


Comment:

- Due to the different slopes produced by the lines connecting the levels, there is a possibility of interaction between these factors. But we need more data to illustrate more meaning of these factors.

```
ggplot(data = trShrew, aes(x = Shrews,
                           y = HeartRates,
                           shape=Sleep,
                           size = Sleep,
                           main = "boxplot of these factors"
                           )) +
  geom_boxplot()
```

```
## Warning: Using size for a discrete variable is not advised.
```



Comment:

- Due to a lack of data in these factors, the box plot with the number of cells is difficult to comprehend.

c. Explain why we cannot fit a Two-Way ANOVA with interaction model to this dataset.

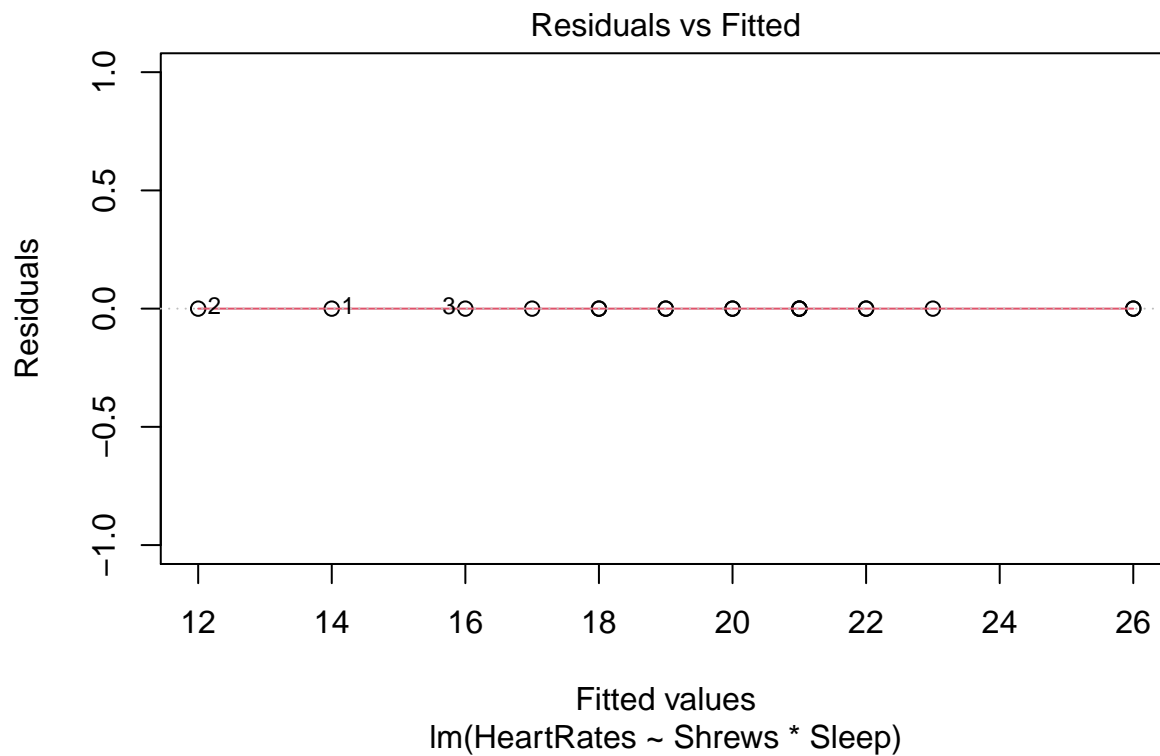
```
trShrew.int = lm(HeartRates ~ Shrews * Sleep, data = trShrew)
anova(trShrew.int)
```

```
## Warning in anova.lm(trShrew.int): ANOVA F-tests on an essentially perfect fit
## are unreliable
```

```
## Analysis of Variance Table
```

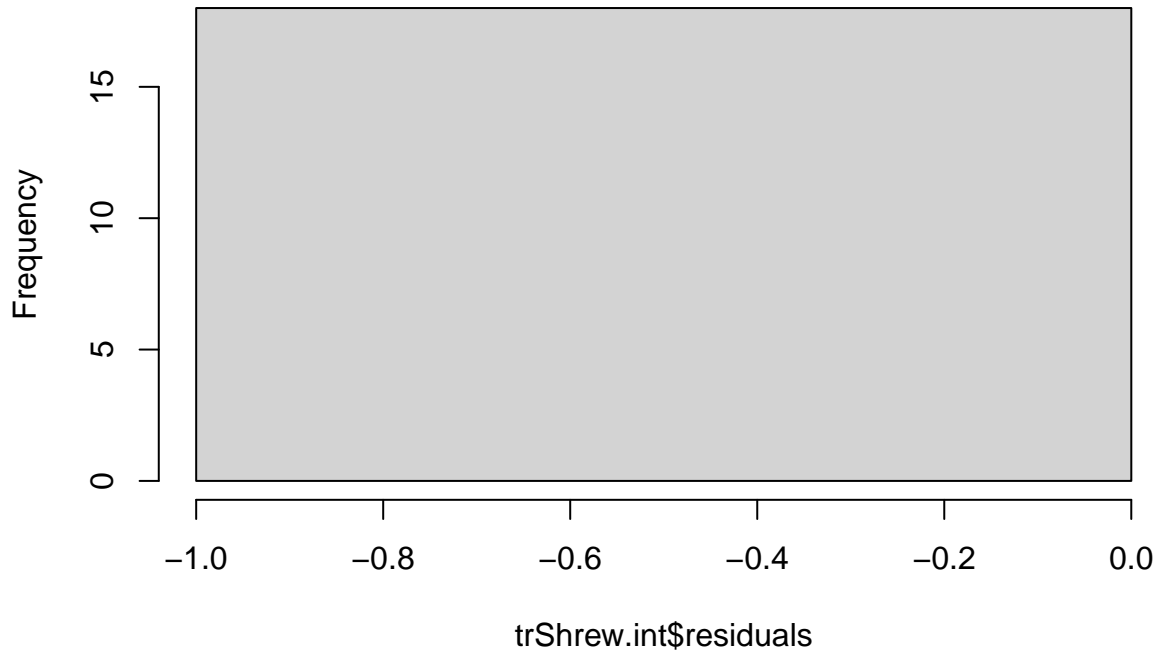
```
##
## Response: HeartRates
##           Df Sum Sq Mean Sq F value Pr(>F)
## Shrews      5 186.278   37.256
## Sleep       2  14.778    7.389
## Shrews:Sleep 10  24.556    2.456
## Residuals   0    0.000
```

```
plot(trShrew.int, which = 1)
```



```
hist(trShrew.int$residuals)
```

Histogram of trShrew.int\$residuals



Comment:

- Since the residual is zero, the interaction model for this data is unsuitable for Two-way Anova. Besides, the residual distribution is not normal distribution.
- Therefore, we remove the interaction of Shrews and Sleeps which is not significant.

```
trShrew.int.2 = update(trShrew.int, . ~ . - Shrews:Sleep)
anova(trShrew.int.2)
```

```
## Analysis of Variance Table
##
## Response: HeartRates
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Shrews      5 186.278   37.256  15.172 0.0002157 ***
## Sleep       2  14.778    7.389   3.009 0.0948298 .
## Residuals  10  24.556    2.456
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Order doesn't matter in Linear regression framework
trShrew.int.3 = update(trShrew.int.2, . ~ . - Sleep)
anova(trShrew.int.3)
```

```
## Analysis of Variance Table
##
## Response: HeartRates
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Shrews      5 186.278   37.256  11.366 0.0003231 ***
```

```
## Residuals 12 39.333 3.278
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment:

- Reviewing the p value of $lm(\text{HeartRate} \sim \text{Shrews} + \text{Sleep})$, we update to remove the Sleep which is insignificant.
- Next, we transform a log model to acquire a better p value.

```
trShrew.int.log = lm(log(HeartRates) ~ Shrews, data = trShrew)
drop1(trShrew.int.log, test = "F")
```

```
## Single term deletions
##
## Model:
## log(HeartRates) ~ Shrews
##          Df Sum of Sq      RSS       AIC F value    Pr(>F)
## <none>                 0.10526 -80.550
## Shrews   5    0.55135 0.65661 -57.599  12.571 0.0001994 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Therefore, we choose the log model with only the Shrews predictor.

d. check assumptions:

d1. Write down the mathematical model for this situation, defining all appropriate parameters

Consider the One-Way ANOVA:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon; \quad \varepsilon \sim N(0, \sigma^2)$$

where:

- Y : HeartRates - The ratio of HeartRates observed.
- X_1 : Shrews variables respectively.
- β_1 : Shrews regression coefficients respectively.
- β_0 : Intercept term

d2. State the appropriate hypotheses

Hypotheses of Anova test in a linear regression:

$$H_0 : \beta_1 = 0; \text{ and } H_1 = \beta_1 \neq 0$$

d3. Compute an appropriate ANOVA table

- The anova will be:

```
anova(trShrew.int.log)
```

```
## Analysis of Variance Table
##
## Response: log(HeartRates)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Shrews      5 0.55135 0.110269  12.571 0.0001994 ***
## Residuals  12 0.10526 0.008772
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

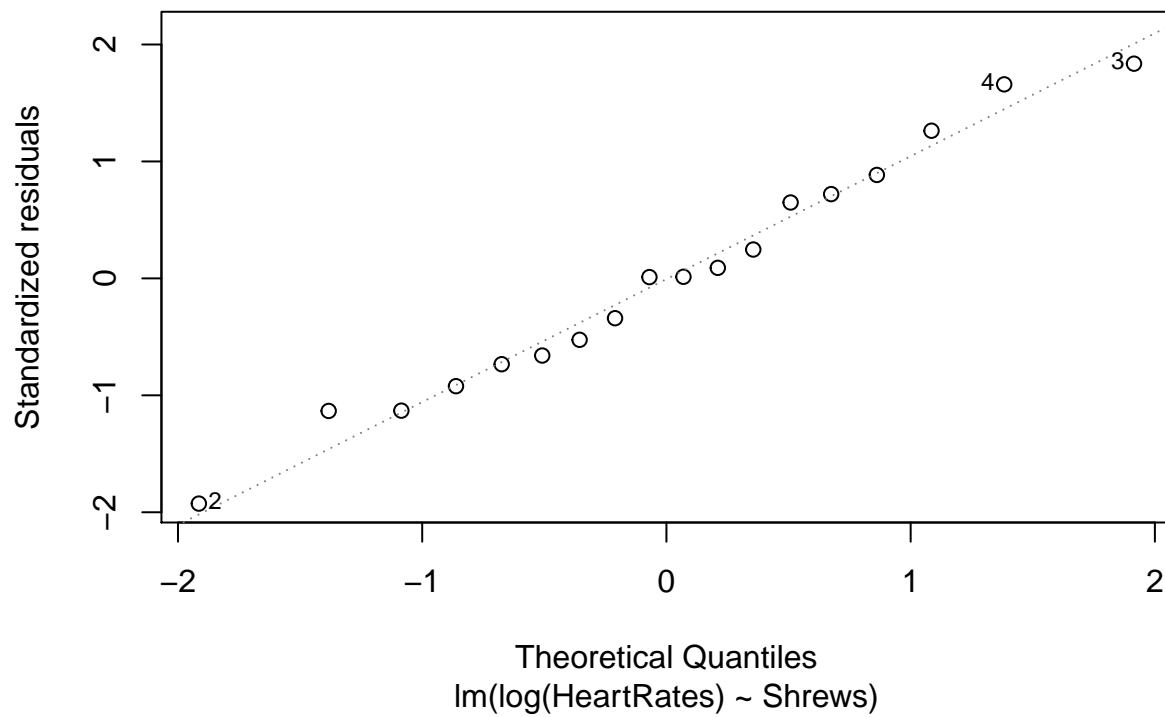
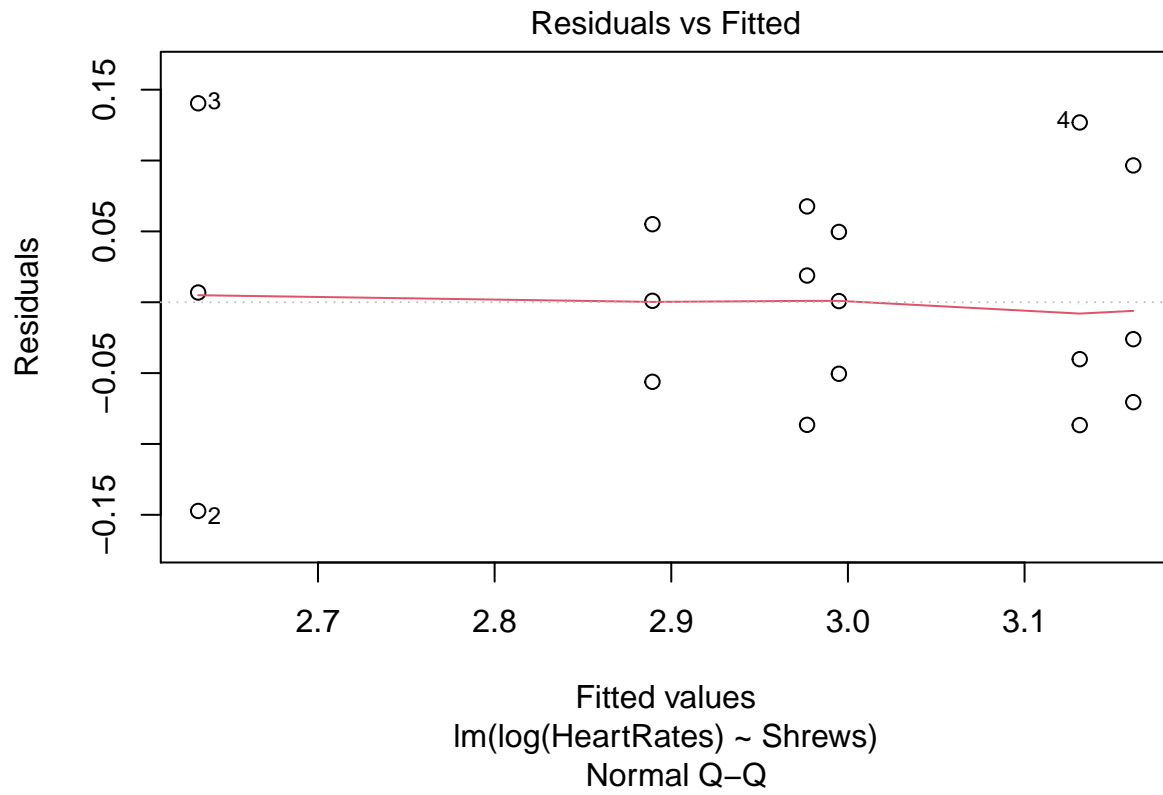
- Summary of model will be:

```
summary(trShrew.int.log)
```

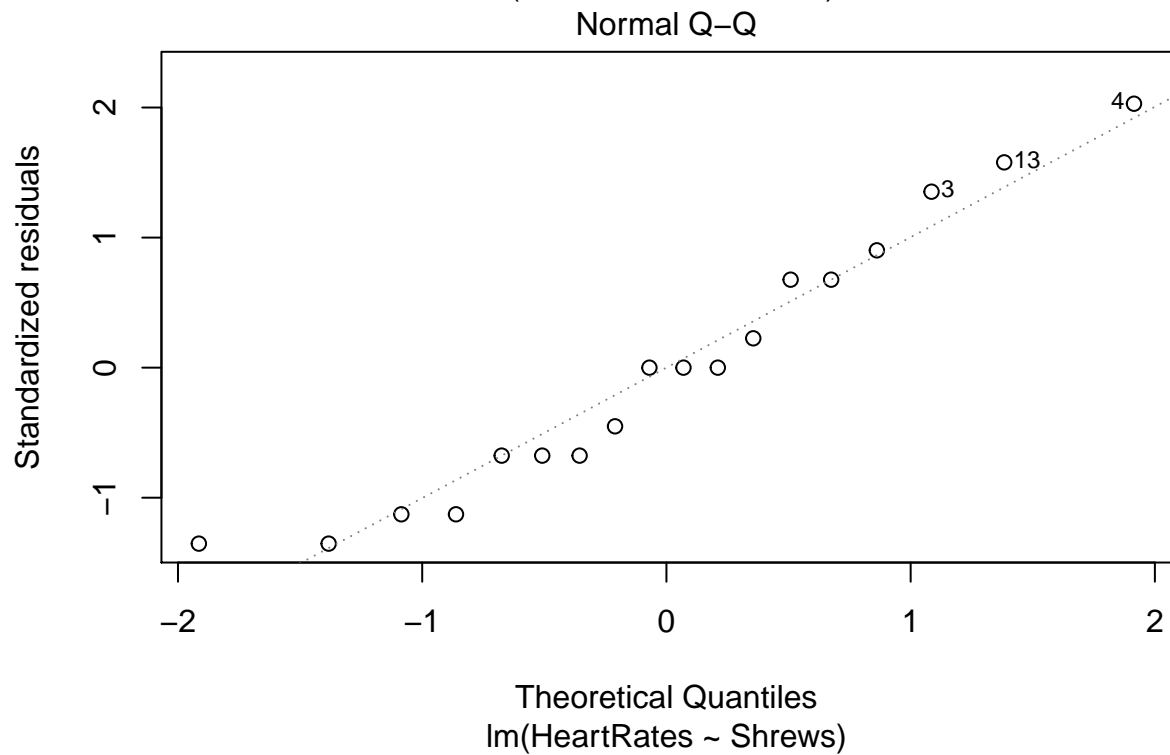
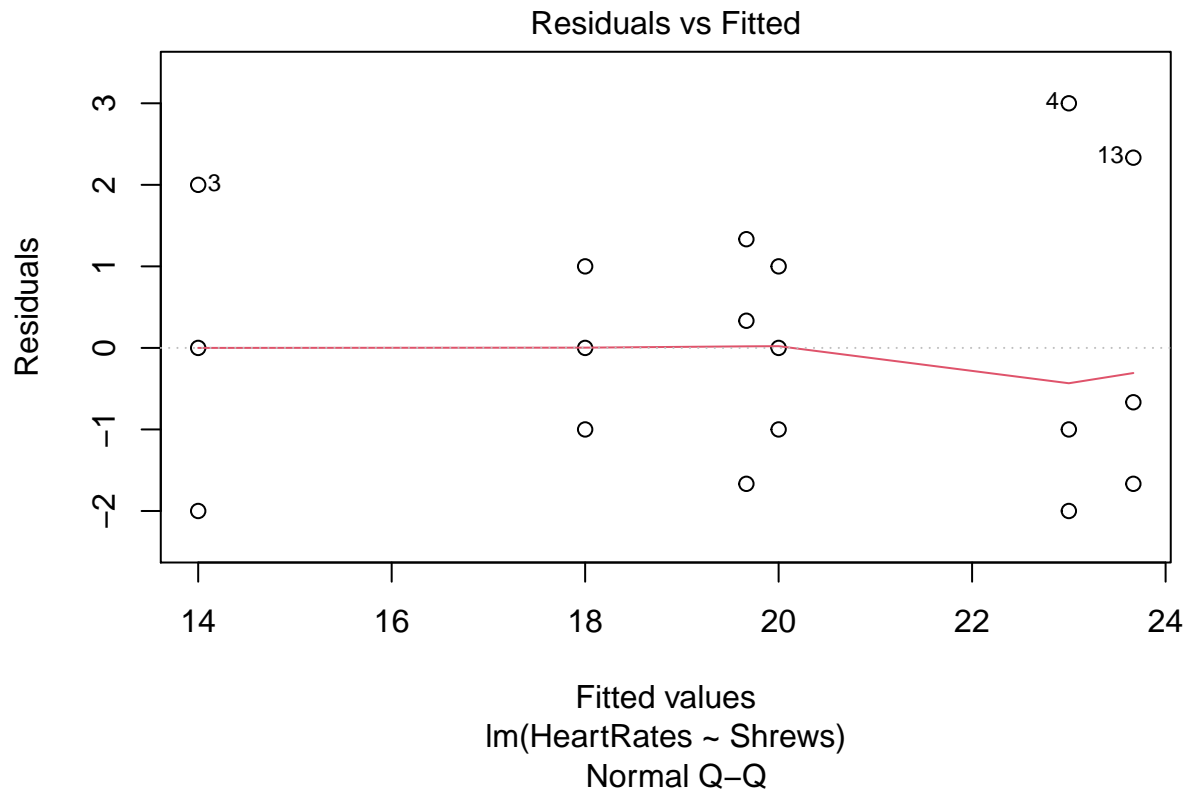
```
##
## Call:
## lm(formula = log(HeartRates) ~ Shrews, data = trShrew)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.147278 -0.054711  0.000932  0.053729  0.140404
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.63218    0.05407  48.678 3.70e-15 ***
## Shrews2      0.49904    0.07647   6.526 2.83e-05 ***
## Shrews3      0.34469    0.07647   4.507 0.000717 ***
## Shrews4      0.25716    0.07647   3.363 0.005645 **
## Shrews5      0.52936    0.07647   6.922 1.60e-05 ***
## Shrews6      0.36271    0.07647   4.743 0.000478 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09366 on 12 degrees of freedom
## Multiple R-squared:  0.8397, Adjusted R-squared:  0.7729
## F-statistic: 12.57 on 5 and 12 DF, p-value: 0.0001994
```

d4. Check assumptions

```
plot(trShrew.int.log, which = 1:2)
```



```
plot(trShrew.int.3, which = 1:2)
```

Comment:

- The normal quantile plot of residuals is closer to linear suggesting residuals are close to normally distributed.
- There is no pattern in the residual plot vs fitted values, the variability seems constant between effects.

- The log model appears slightly better than it does without the log model.

e. State your conclusions about the effect of Shrews and Sleep on HeartRates. You do not need to statistically examine the multiple comparisons.

- According to the findings, the Shrew factor has a significant effect on the number of HeartRates. While the Sleep factor has a insignificant effect on HeartRates.