Assignment2 - STAT8121

Minh Tien Ta - 46207031

1 Oct 2021

Contents

Question1

Reading dataset

```
paramo <- read.table('data/paramo.dat',header=T)
print(dim(paramo))
```

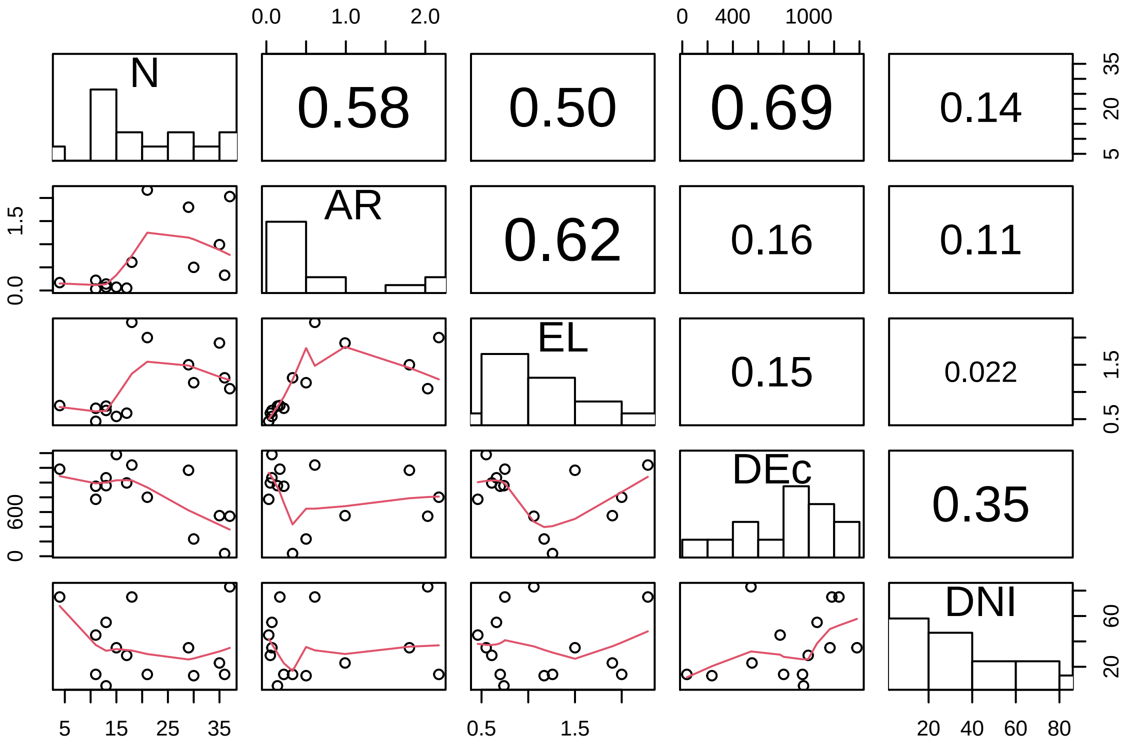
```
## [1] 14  5
```

```
tail(paramo,3)
```

```
##     N   AR   EL  DEc DNI
## 12  4 0.17 0.75 1182  75
## 13 18 0.61 2.28 1238  75
## 14 15 0.07 0.55 1380  35
```

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,
      diag.panel  = panel.hist,
      lower.panel = panel.smooth,
      # pch = "."
      )
```



Comments:

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.
- X1,X2,X3,X4 : 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:

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

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

```
n = nrow(paramo) ; k  = ncol(paramo) ; n ; k
```

```
## [1] 14
```

```
## [1] 5
```

```
df1 = k - 1; df2 = n - k; df1 ;df2
```

```
## [1] 4
```

```
## [1] 9
```

```
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

If $p_{value} \leq \alpha$ reject the null hypothesis. $p_{value} > \alpha$ If fail to reject the null hypothesis

b6. Compute the P-Value

we can caculate that the 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)

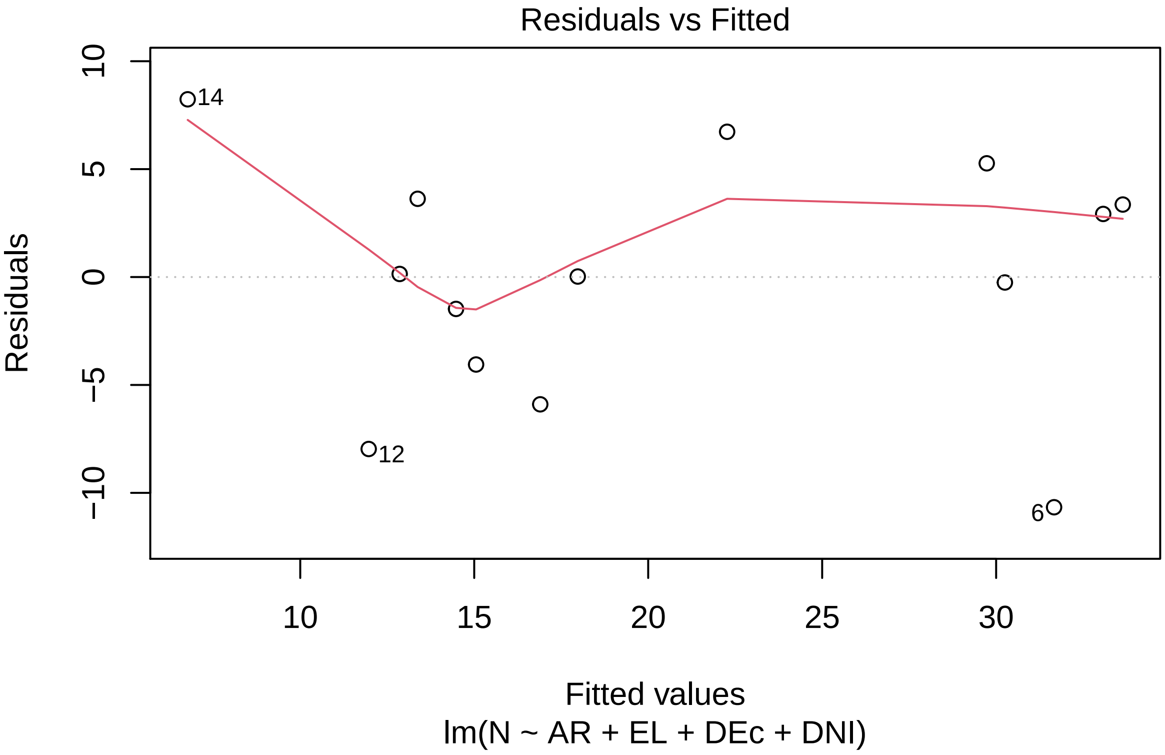
There is a significant linear relationship between percentage response N and at least one of the four predictor variables.

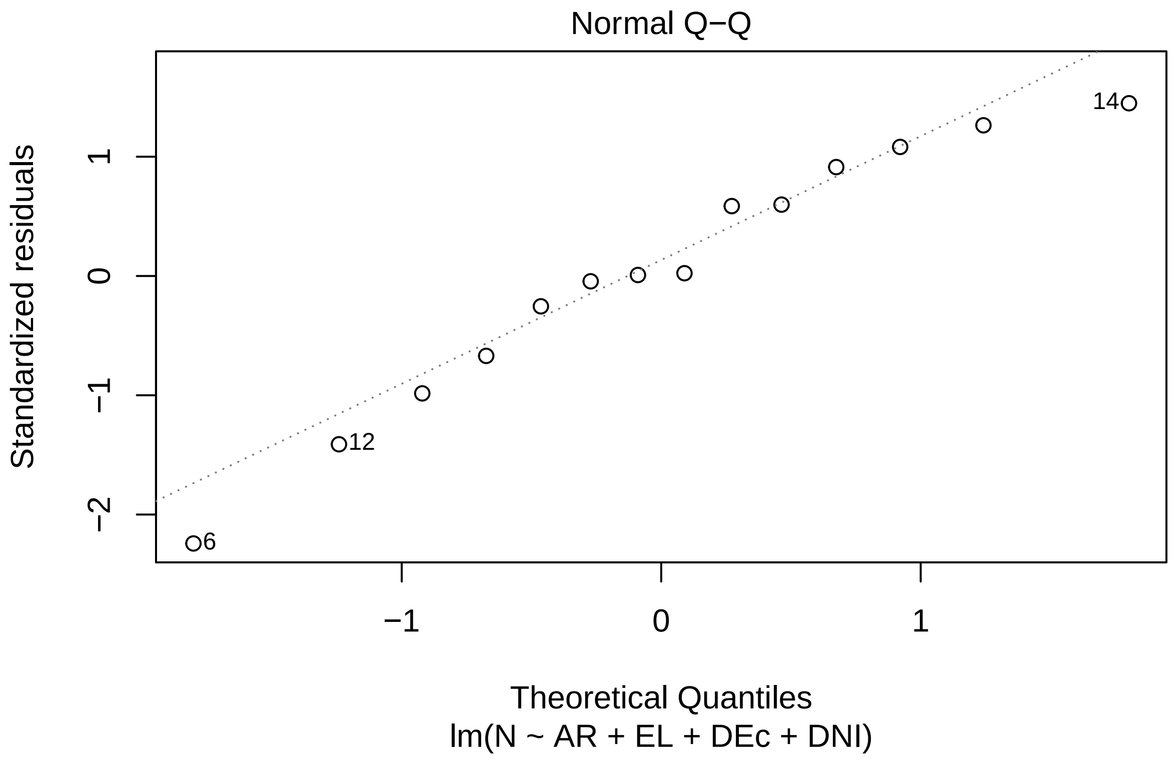
c. Validate the full model using all the predictors and comment on whether it is appropriate to a multiple regression model to explain the N abundance value.

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

Check diagnostics:

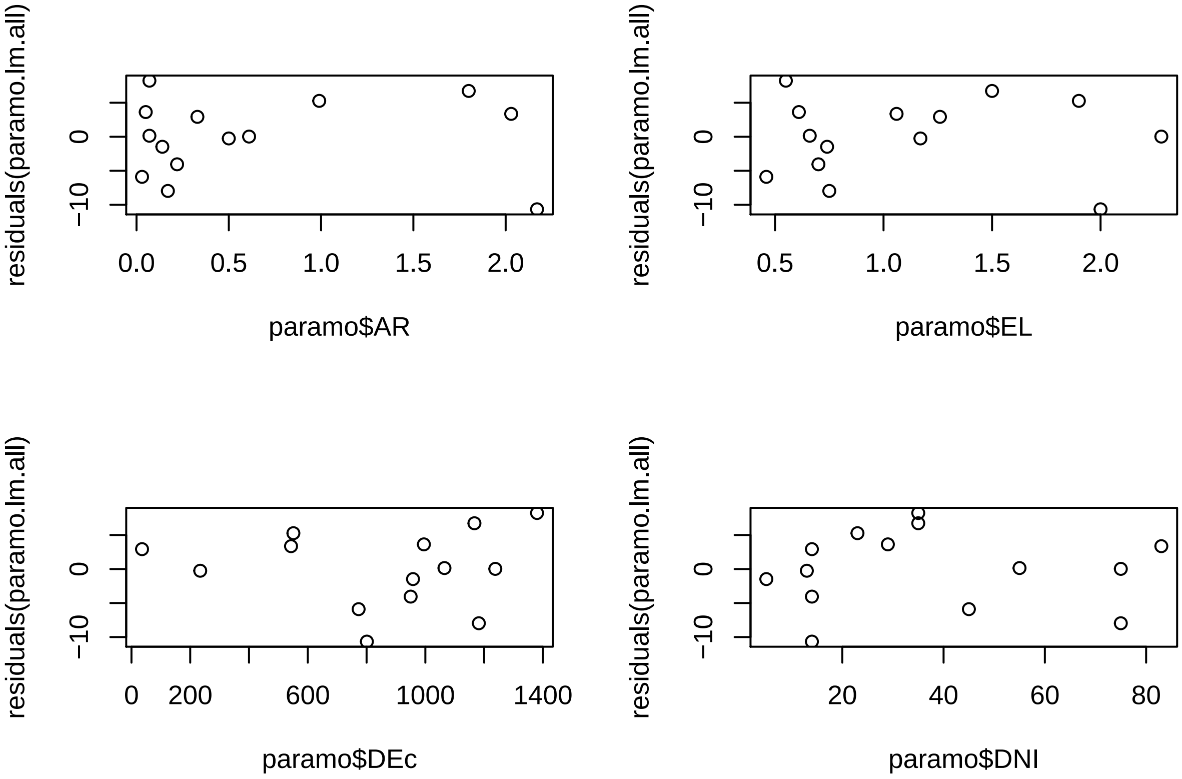
```
plot(paramo.lm.all, which = 1:2)
```





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))
```



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:

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

```
total_SumSQ = sum(paramo.aov[["Sum Sq"]][1:5]); total_SumSQ
```

```
## [1] 1498.857
```

```
SQ_reg = paramo.aov[["Sum Sq"]][5] ; SQ_reg
```

```
## [1] 404.5895
```

```
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 insignficant variables like DNE

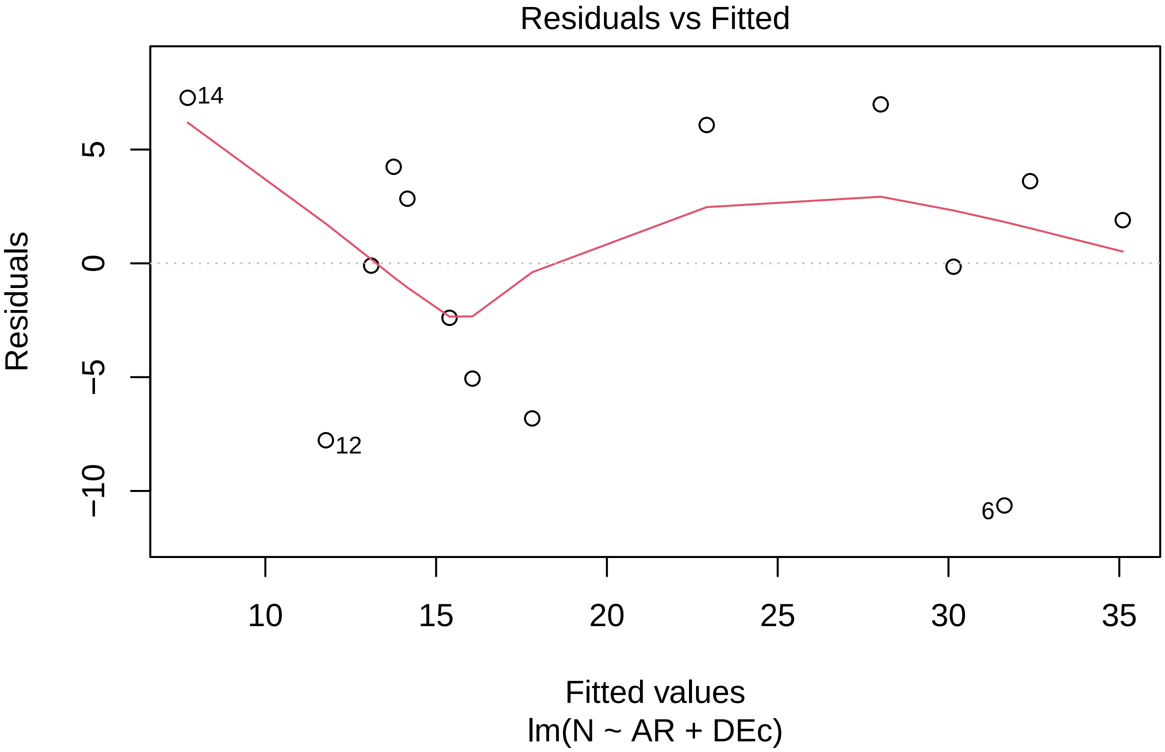
```
para.lm.2 = lm(formula = N ~ AR + EL + DEc , data = paramo)
summary(para.lm.2)
```

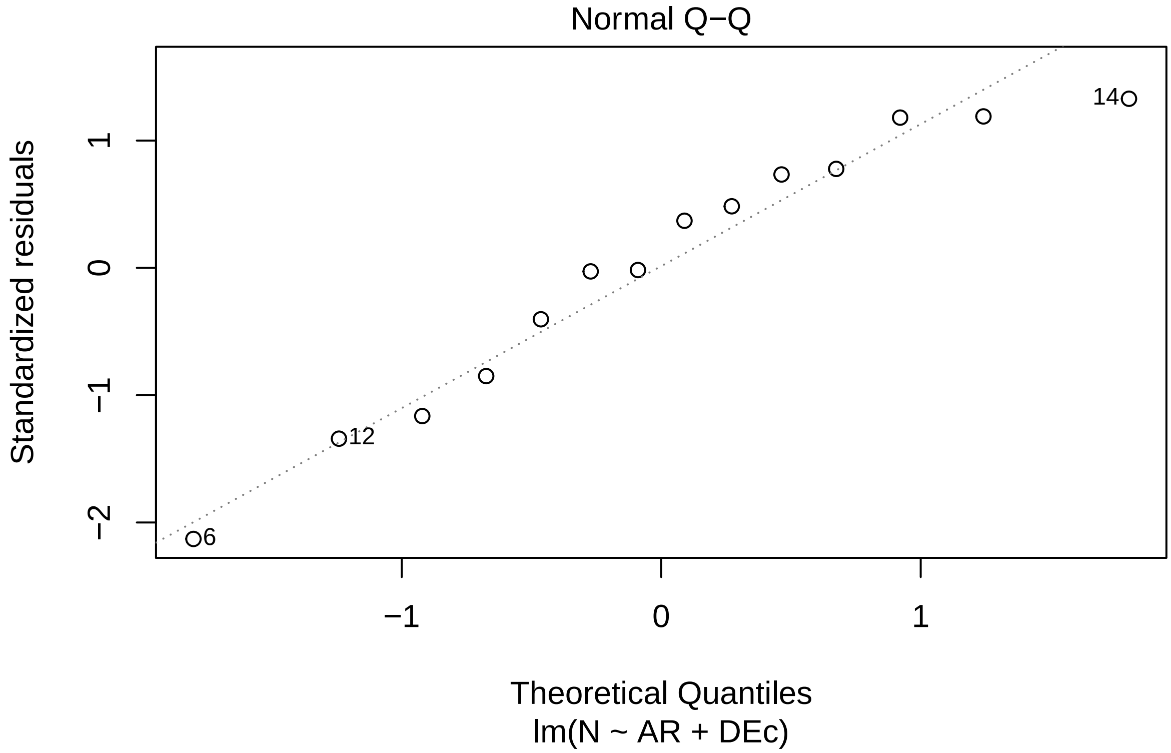
```
##
## Call:
## lm(formula = N ~ AR + EL + DEc, data = paramo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1638  -3.8306   0.4693   3.9477   8.0285
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.10415    5.80141   4.844 0.000677 ***
## 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
```

```
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)
```





f. Comment on the R2 and Adjusted R2 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 is:

Adjusted $R^2 = R^2 - (1 - R^2)\frac{p-1}{n-p}$ where n = 14 and p = 3 for model 3.

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

```
## Analysis of Variance Table
##
## Response: N
##           Df Sum Sq Mean Sq F value   Pr(>F)
## AR         1 508.92  508.92  12.937 0.004193 **
## DEc        1 557.23  557.23  14.165 0.003134 **
## Residuals 11 432.71   39.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R_2_model3 = (sum(para.anova.3$`Sum Sq`[1:3]) - para.anova.3$`Sum Sq`[3] ) / sum(para.anova.3$`Sum Sq`[
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] ; coef_AR
```

```
## [1] 6.683038
```

```
se_AR = summary(para.lm.3)$coef[5] ; se_AR
```

```
## [1] 2.264403
```

```
alpha = 0.05; t_table = qt(1- alpha/2,11) ; t_table
```

```
## [1] 2.200985
```

```
c(coef_AR - t_table*se_AR,coef_AR + t_table*se_AR)
```

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

Question2

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

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

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

d. check assumptions:

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

d2. State the appropriate hypotheses

d3. Compute an appropriate ANOVA table

d4. Check assumptions

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