Assignment_Anthony_Carino_46375821

Anthony Carino 46375821

17/10/2021

Question 1

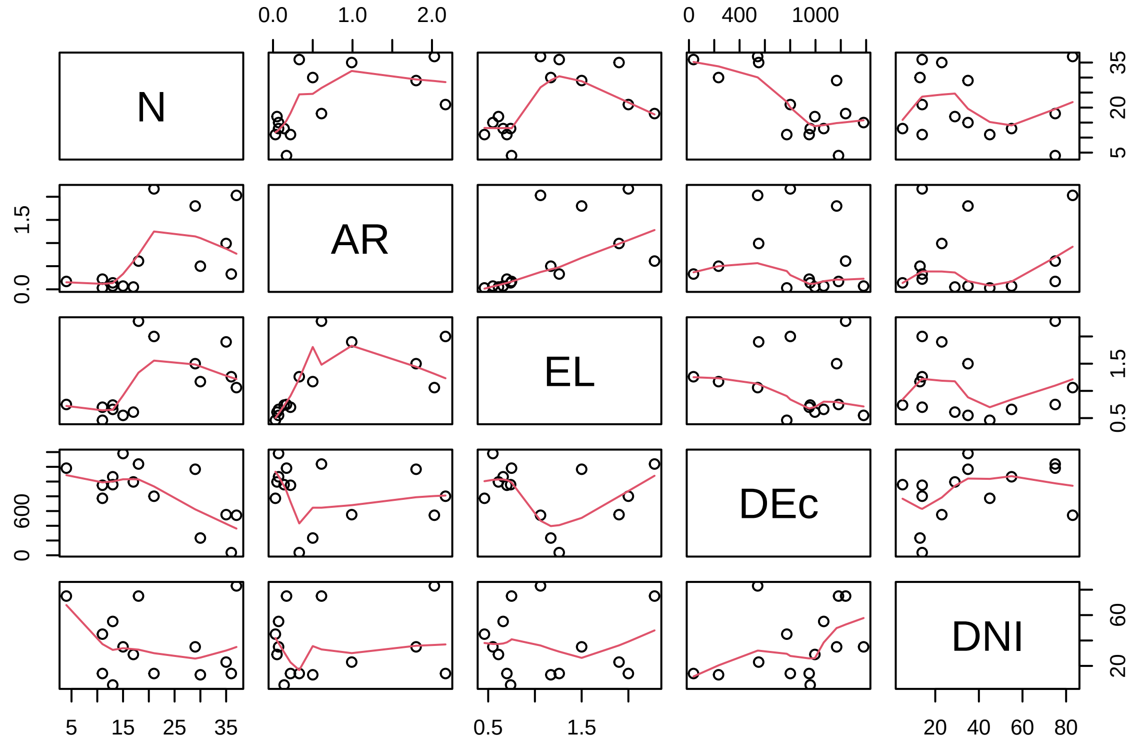
a) These outputs help to identify possible predictors to the N response variable. The correlation matrix shows there is a potentially significant positive relationship between the response variable N, and AR predictor. They have a 0.583 correlation. The response variable N also has a potential strong negative relationship with the DEc predictor with a -0.695 correlation.

These potentially significant relationships are visually highlighted in the scatter plot where the corresponding predictor plots demonstrate stronger relationships with the response variable. However, this is not conclusive. These relationships need to be tested.

```
cor(paramo)
```

```
##              N         AR          EL        DEc         DNI
## N    1.0000000  0.5826995  0.49836214 -0.6947685 -0.13507551
## AR   0.5826995  1.0000000  0.61951650 -0.1593048  0.11159147
## EL   0.4983621  0.6195165  1.00000000 -0.1539371  0.02179708
## DEc -0.6947685 -0.1593048 -0.15393710  1.0000000  0.35416304
## DNI -0.1350755  0.1115915  0.02179708  0.3541630  1.00000000
```

```
pairs(paramo, panel = panel.smooth)
```



b) Fit model using all predictors to explain N.

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

```
##
## Call:
## lm(formula = N ~ AR + EL + DEc + DNI, data = paramo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6660  -3.4090   0.0834   3.5592   8.2357
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.889386   6.181843   4.511  0.00146 **
## AR           5.153864   3.098074   1.664  0.13056
## EL           3.075136   4.000326   0.769  0.46175
## DEc         -0.017216   0.005243  -3.284  0.00947 **
## DNI          0.016591   0.077573   0.214  0.83541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.705 on 9 degrees of freedom
## Multiple R-squared:  0.7301, Adjusted R-squared:  0.6101
## F-statistic: 6.085 on 4 and 9 DF,  p-value: 0.01182
```

Mathematical Multiple Regression Model:

$\hat{N} = \beta_0 + \beta_1(AR) + \beta_2(EL) + \beta_3(DEc) + \beta_4(DNI)$

$\hat{N}$ = 27.889 + 5.154(AR) + 3.075EL - 0.0127(DEc) + 0.0166(DNI)

- This model minimises the residual mean squared and the variation around the regression line (6.705) shown in the output.

Hypothesis:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1: not all $\beta_i = 0$

- We are testing whether the each predictor has a significant impact on the response variable. If the null hypothesis is true, then the predictors do not have an impact on the response.

ANOVA Table:

```
anova(paramo.lm)
```

```
## 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
```

F-Statistic / Null Distribution:

$\textit{Regression Sum of Squares} = 508.92 + 45.90 + 537.39 + 2.06 = 1094.27$

$\textit{Regression Mean Squared} = \frac{Reg.S.S.}{k} = \frac{1094.27}{4} = 273.568$

$\textit{Residual Mean Squared} = 44.95$

$F_{obs} = F_{4,9} = \frac{Reg.M.S.}{Res.M.S.} = \frac{273.568}{44.95} = 6.086$ with 4 and 9 degrees of freedom

- **NOTE:** this can be found in the *summary(paramo.lm)* output.

P-Value:

- $P(F_{3,10} \geq 6.086) = 0.01182$

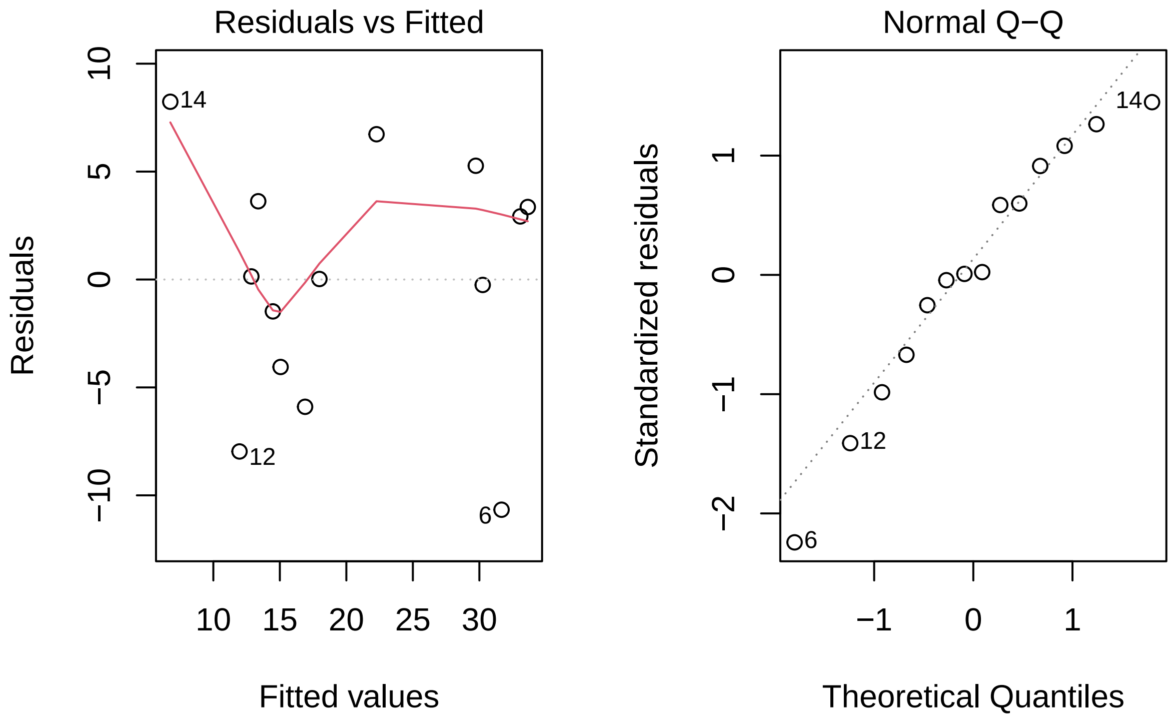
Conclusion:

Since the P-value $0.01182 < 0.05$, the result is significant. There is evidence that the predictors have a significant impact on the response variable.

c) Validate the full model using all the predictors.

The Normal QQ plot is approximately linear with slight skewness, and the Residuals vs Fitted plot shows similar variances. The patterns and skewness can be put down to randomness due to the relatively low number of observations of Tree Shrews. Therefore, the multiple regression model can be used to explain the N abundance value. However, for more accuracy and confidence, the study could include more observations.

```
par(mfrow = c(1,2))
plot(paramo.lm, which = 1:2)
```



d) Find R^2 and comment on what it means in the context of the data set.

R^2 is 0.7301 (found in the output of *summary(paramo.lm)*). This is the coefficient of determination and is the percentage of variation in N that is explained by the linear regression of all the predictors. In this case, 73.01% of the variation in N is explained by the predictors. Given that the rule of thumb is 70%, the model is strong. From the other perspective, approximately 18% of the model was not explained by the predictors and is caused by random variation.

e) Find the best multiple regression model that explains the data.

Using the Step-wise Backward Estimation method, removing the predictor with the highest p-value (the most insignificant) until they are all significant will give the best model. To begin, removing the DNI predictor since it has the highest p-value (0.83541) we have the model:

```
paramo.lm2 = lm(N~AR+EL+DEc,data=paramo)
summary(paramo.lm2)
```

```
##
## 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
```

- Now the EL predictor is the most insignificant so we remove it and refit the model:

```
paramo.lm3 = lm(N~AR+DEc, data=paramo)
summary(paramo.lm3)
```

```
##
## 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
```

- After removing DNI and EL, all other predictors are now significant. therefore we can conclude this is the best multiple regression model to explain the data. Fitting AR and DEc to explain the N abundance.

f) The R^2 and **Adjusted** R^2 measures are different when comparing the 'Best' model and the model containing all predictors. Here in the 'Best' model, the R^2 is 0.7113 which is lower than the model containing all predictors (0.7301). This is because R^2 will get higher as you include more predictors to fit the model, even if they are insignificant.

The **Adjusted** R^2 rectifies this by penalising the extra predictors. Now, the **Adjusted** R^2 for the 'Best' model is higher (0.6588) than the model with all predictors (0.6101).

g) Compute 95% confidence interval for the AR regression parameter.

```
## [1] 2.160369
```

- 95% CI = estimate +/- critical value * standard error
- 95% CI for AR = 6.683 +/- 2.160369 * 2.2644
  (1.7912, 11.5749)

We are 95% confident that each unit increase in N will cause the increase in AR to be between (1.7912, 11.5749). Given that this interval doesn't contain 0, it highlights how the effect of AR is significant in predicting the N response variable.

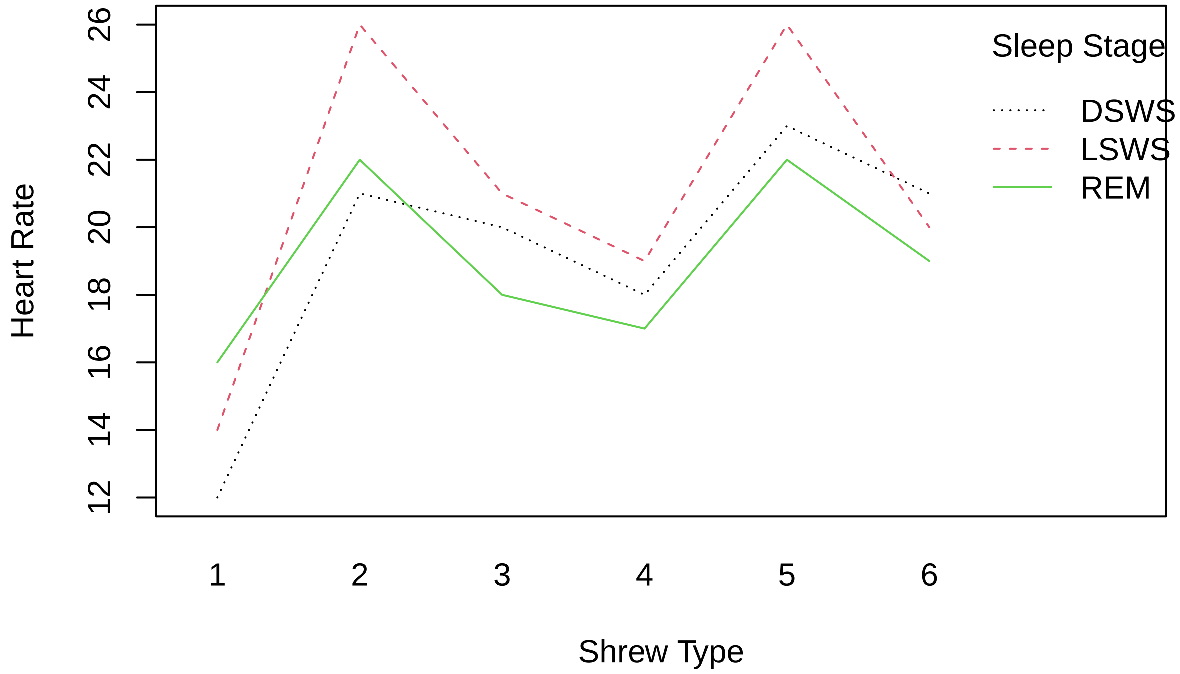
Question 2

```
shrews = read.table("TreeShrews.dat", header = TRUE)
```

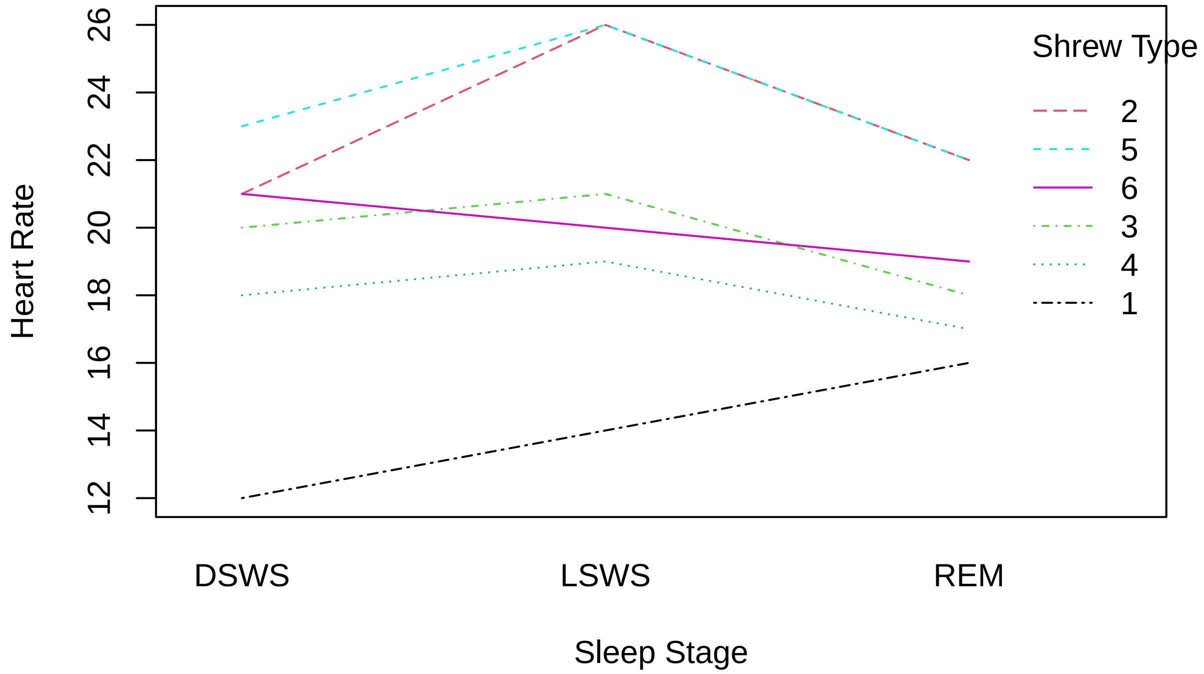
a) This study is a balanced design because there are the same number of replicates for each treatment combination. There are three replicates for each type of Tree Shrew, as well as one of each sleep stage for each type of Shrew.

b) Interaction Plots

```
#Plot 1
interaction.plot(shrews$Shrews, shrews$Sleep, shrews$HeartRates, trace.label = "Sleep Stage", xlab = "Sl
```



```
#Plot 2
interaction.plot(shrews$Sleep, shrews$Shrews, shrews$HeartRates, trace.label = "Shrew Type", xlab = "Sl
```



- Comment: some of the lines are somewhat parallel graph meaning that they do not have significant effects on the response variable. For example, in Plot 2, Shrew type 3,4 and 5 are parallel and do not cross. In Plot 1, there is a possible interaction effect between the sleep stage given there is a cross over between lines.

c) We cannot fit a Two-Way ANOVA with interaction model to this data set because the interaction effect of Shrews and Sleep is 0.727, which is insignificant. Therefore, it must be removed and test the main effects individually.

```
shrewsANOVA = lm(HeartRates~Shrews*factor(Sleep),data=shrews)
anova(shrewsANOVA)
```

```
## Analysis of Variance Table
##
## Response: HeartRates
##                      Df  Sum Sq Mean Sq F value Pr(>F)
## Shrews                1  39.433  39.433  2.9114 0.1137
## factor(Sleep)         2  14.778   7.389  0.5455 0.5933
## Shrews:factor(Sleep)  2   8.867   4.433  0.3273 0.7271
## Residuals            12 162.533  13.544
```

d) Carrying out the Individual Tests

Test 1: Fitting the ‘Sleep’ effect on ‘HeartRates’

```
sleep.lm = lm(HeartRates~factor(Sleep),data=shrews)
summary(sleep.lm)
```

```
##
## Call:
## lm(formula = HeartRates ~ factor(Sleep), data = shrews)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.167 -1.792  0.000  2.708  5.000
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        19.1667     1.5306  12.523 2.41e-09 ***
## factor(Sleep)LSWS   1.8333     2.1645   0.847     0.41
## factor(Sleep)REM   -0.1667     2.1645  -0.077     0.94
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.749 on 15 degrees of freedom
## Multiple R-squared:  0.0655, Adjusted R-squared:  -0.0591
## F-statistic: 0.5257 on 2 and 15 DF,  p-value: 0.6016
```

Mathematical Model:

- HeartRates $=$ 19.1667 $+$ 1.8333(LSWS) - 0.1667(REM)

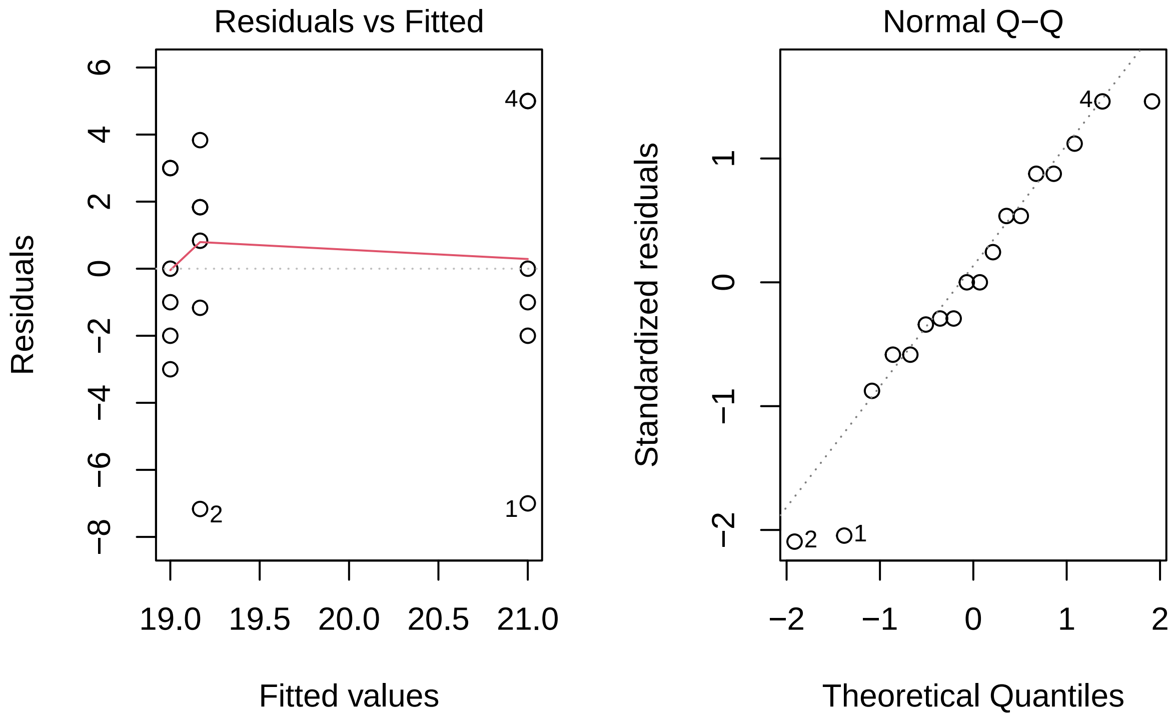
Hypothesis:

- H_0: $\beta_i = 0$ (all Sleep stages are the same); H_1: not all β are the same

```
anova(sleep.lm)
```

```
## Analysis of Variance Table
##
## Response: HeartRates
##               Df  Sum Sq Mean Sq F value Pr(>F)
## factor(Sleep)  2  14.778  7.3889  0.5257 0.6016
## Residuals     15 210.833 14.0556
```

```
#assumptions for sleep.lm
par(mfrow = c(1,2))
plot(sleep.lm,which=1:2)
```



- The assumption of equal variance is upheld as shown in the Residuals vs Fitted plot. However, the Normal QQ plot suggests that there is a slight skewness in the distribution of residuals. The results of this model should be assessed with caution based on this.

Test 2: Fitting the 'Shrews' effect on 'HeartRates'

```
shrews.lm = lm(HeartRates~factor(Shrews),data=shrews)
summary(shrews.lm)
```

```
##
## Call:
## lm(formula = HeartRates ~ factor(Shrews), data = shrews)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -2     -1      0      1      3
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       14.000      1.045  13.394 1.41e-08 ***
## factor(Shrews)2    9.000      1.478   6.088 5.43e-05 ***
## factor(Shrews)3    5.667      1.478   3.833  0.00238 **
## factor(Shrews)4    4.000      1.478   2.706  0.01910 *
## factor(Shrews)5    9.667      1.478   6.539 2.77e-05 ***
## factor(Shrews)6    6.000      1.478   4.059  0.00158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.81 on 12 degrees of freedom
## Multiple R-squared:  0.8257, Adjusted R-squared:  0.753
## F-statistic: 11.37 on 5 and 12 DF,  p-value: 0.0003231
```

- *Mathematical Model:*

  $\text{HeartRates} = 14 + 9(\text{Shrews2}) + 5.667(\text{Shrews3}) + 4(\text{Shrews4}) + 9.667(\text{Shrews5}) + 6(\text{Shrews6})$.

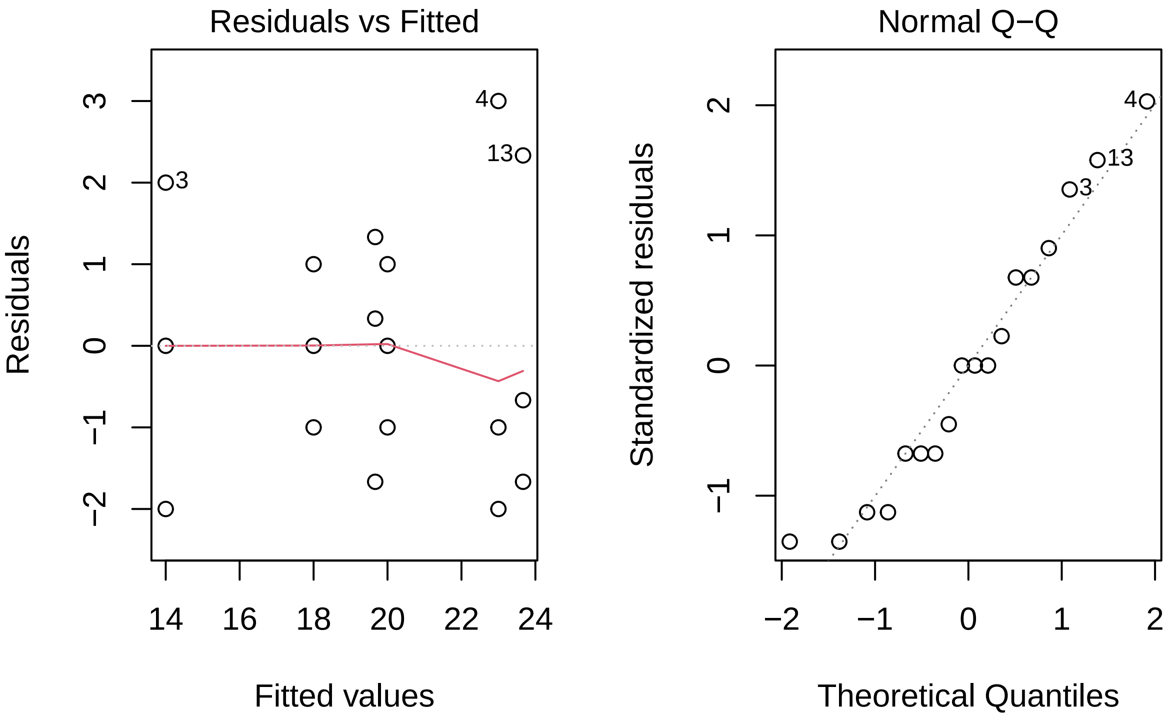
- *Hypothesis:*

  H_0: $\beta_i = 0$ (all shrew types are the same); H_1: not all β are the same

```
anova(shrews.lm)
```

```
## Analysis of Variance Table
##
## Response: HeartRates
##                 Df  Sum Sq Mean Sq F value    Pr(>F)
## factor(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
```

```
#assumptions for shrews.lm
par(mfrow = c(1,2))
plot(shrews.lm,which=1:2)
```



- The Residuals vs Fitted plot shows no significant pattern and equal variances. The Normal QQ plot also suggests that the distribution of residuals are normal. The slight curve and patterns in both plots can be caused by the relatively low number of observations per treatment. Therefore, the assumptions have been satisfied.

e) Conclusions

Since the interaction effect was insignificant, the Sleep and Shrew type predictors were tested individually to see if they had a significant impact on Heart Rates. The Sleep effect did not have any significant impact on Heart Rates because its p-value (0.6016) is greater than 0.05. Therefore it is not a good predictor. However, the Shrew type predictor was a significant predictor of shrew Heart Rates, since its p-value $(0.0003231) < 0.05$. The fitted model has an $R^2 = 0.8257$, meaning that over 82% of the change in Heart Rate is explained by Shrew type. This significance is supported by the first preliminary graph which shows large changes in heart rates for every Shrew type.