STAT2170_Assignment

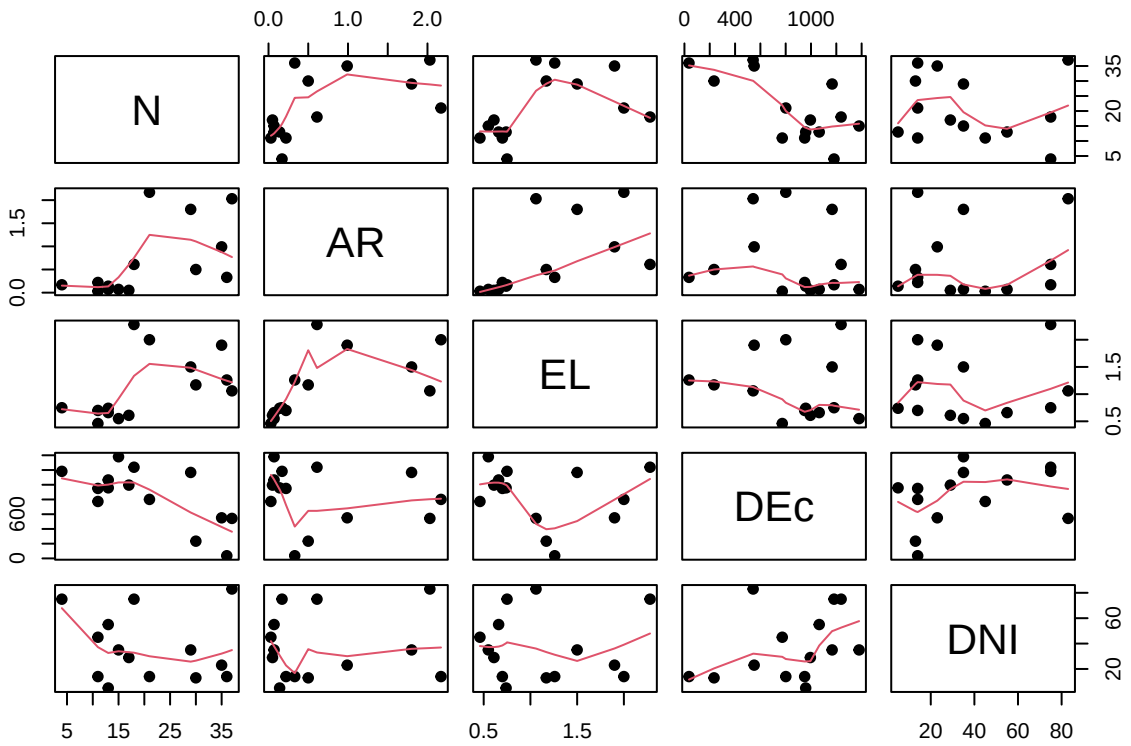
Lachlan Dixon 46478116

21/10/2021

Question 1

a) Creating Scatterplot and Correlation Matrix.

```
pairs(paramo,panel=panel.smooth,pch=19);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
```

There appears to be a positive relationship between N & AR, supported by a moderate 0.58 correlation coefficient, there is a moderately weak relation between N & EL with a correlation coefficient of 0.50, both N & AR and N & EL appear to cluster at small values and spread as they increase. N & DEc has a negative linear relation with a strong correlation coefficient of -0.69. For the explanatory variables, AR & EL appear to have a positive relation albeit with an increasing variance, correlation of 0.62 not high enough for multi-collinearity which would need to be closer to 1. All other predictor variables have weak if any relation.

b) Fit a model using all predictors, Test for overall significance.

$$Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + u$$
$$N = AR + EL + BEc + DNI$$

N: Number of species of bird present
AR: Area of the island in square km
EL: Elevation in thousands of metres
DEc: Distance from ecuador in kilometres
DNI: Distance to the nearest other island in kilometre

Hypothesis:

$H_0 : \beta_{AR} = \beta_{EL} = \beta_{BEc} = \beta_{DNI}$ *against,* $H_1 : \beta_i \neq 0$ *for at least one* i

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

ANOVA Significance:

```
## 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
```

$FullRegss = RegSS_{AR} + RegSS_{EL|AR} + RegSS_{BEc|AR,BEc} + RegSS_{DNI|AR,EL,BEc}$

Full Regss = 508.92+45.90+537.39 +2.06 = 1094.27 = $\frac{1094.27}{K}$ = $\frac{1094.27}{4}$=273.5675

Test-Statistic:

$Fobs = \frac{Reg.M.S}{RESs.M.S} = \frac{273.5675}{44.95} = 6.086$

P-value $= P(F_{4,9} \geq 6.086) = 0.0118166$

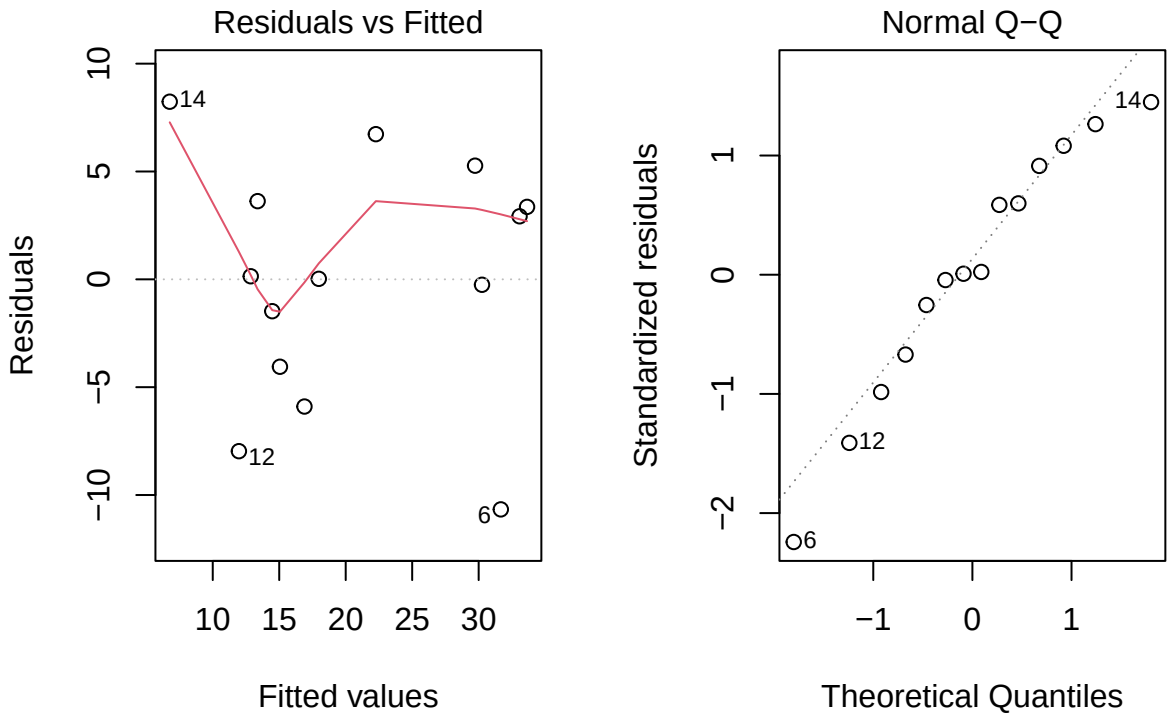
Conclusion:

At significance level of 5% we reject the null hypothesis and conclude that at least one of the parameters do not equal zero.

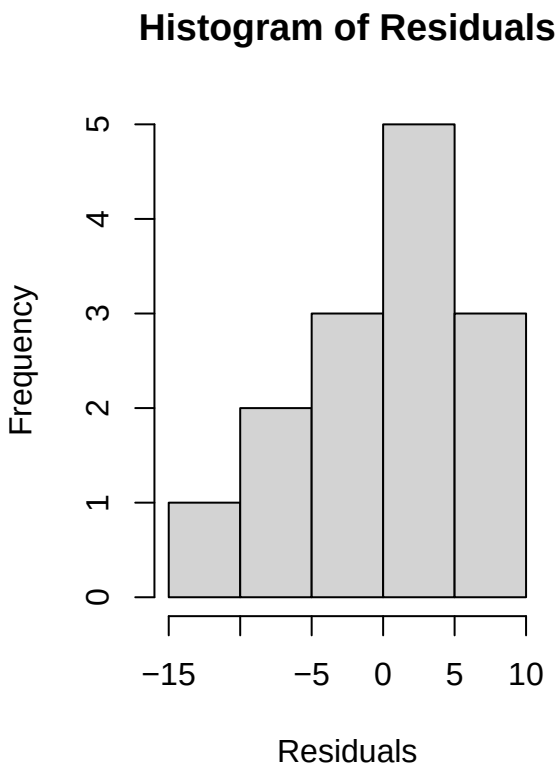
c) Validating the model with all predictors, is it appropriate for the multiple regression model to explain the N abundance value?

Need to first assess residuals with normal Q-Q plot to assume normality, and equal variances.

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




```
hist(lm.paramo$residuals, main = "Histogram of Residuals", xlab = "Residuals")
```



Residuals vs fitted appears to move below then above the zero value casting doubt over randomness of variance, the normal Q-Q plot appears roughly linear satisfying normality. Influencing our decision may be the limited amount of observations, ideally we would like more observations to make a more confident statement about the validation of the model. Histogram of residuals also appears to show a slight left skew. Due to the low observations levels it may be valid to proceed with regression.

d) Find Coefficient of Determination and comment.

Coefficient of Determination $R^2 = \frac{Explained\,Variation}{Total\,Variation} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$

From ANOVA:
SST= 508.92+45.90+537.39+2.06+404.59=**1498.86**
SSR=**404.59**
SSE=SST-SSR = 1498.86-404.59=**1094.27**

$R^2 = \frac{1094.27}{1498.86} = 0.7301$

The proportion of within-sample variation in the dependent variable explained by the model has a range of $1 \geq R^2 \geq 0$. The model provided acccounts for 0.73 of the variance leaving 0.27 unexplained and missing from the model.

e) Using model selection procedures, find the best multiple regression model.

```
summary(lm.paramo)
```

```
##
## 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
```

Can see at 1% significance AR,EL and DNI are insignificant in explaining the model.

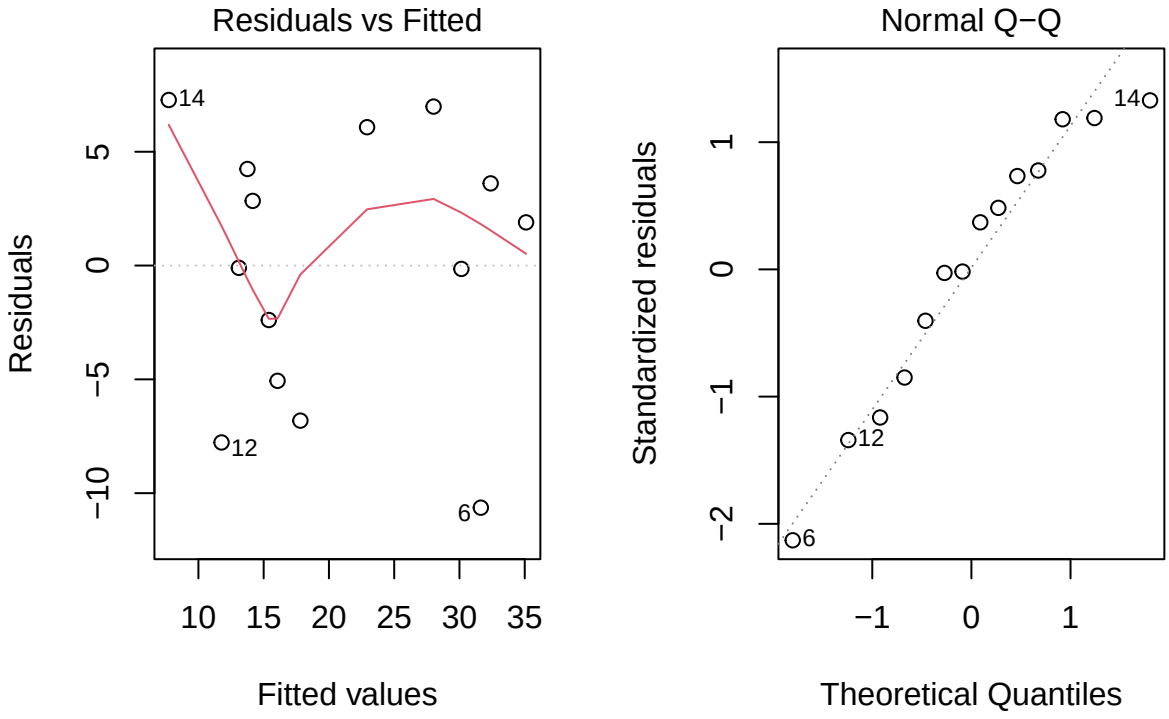
Dropping DNI from the model:

```
lm.paramo.new<-lm(N~AR+DEc,data=paramo)
summary(lm.paramo.new)
```

```
##
## 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
```

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



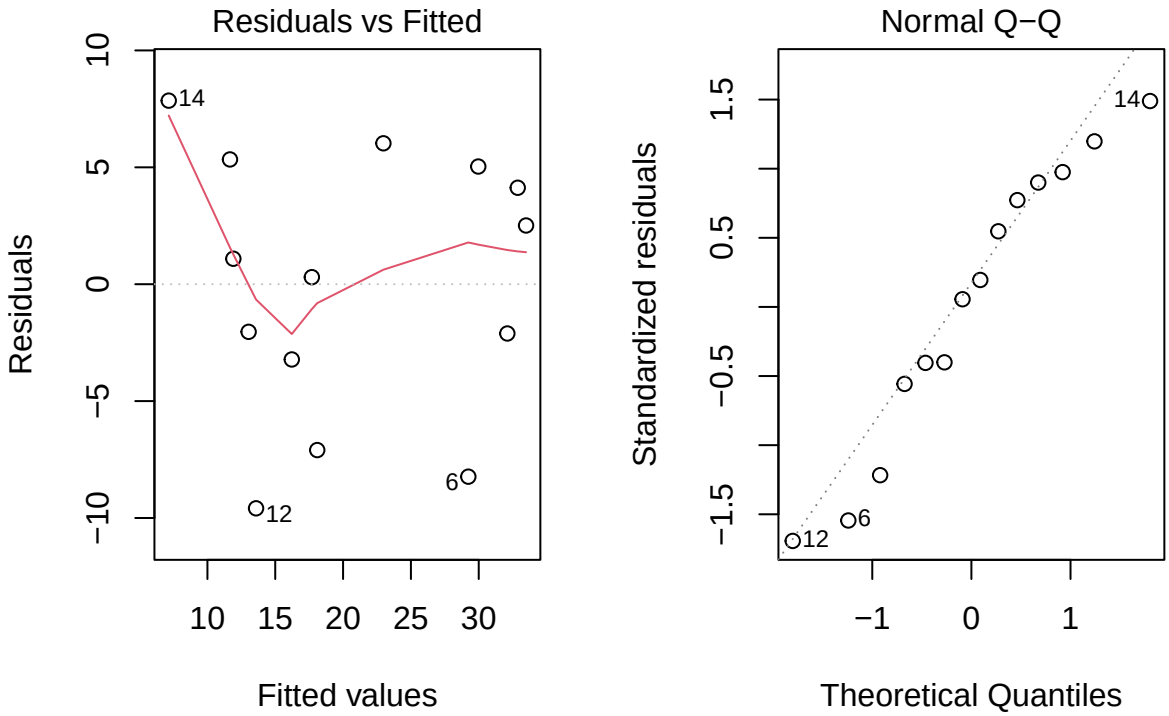
Can see new model appears to improve values, all coefficients being highly significant. Residuals against fitted seems to still show signs of non random variability. N & AR in pairwise appear to show signs of clustering at small values and has a distribution similar to a log model. Taking log of AR:

```
lm.paramo.Final<-lm(N~log(AR)+DEc,data=paramo)
summary(lm.paramo.Final)
```

```
##
## Call:
## lm(formula = N ~ log(AR) + DEc, data = paramo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5860 -2.9414  0.6952  4.8068  7.8520
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.336424   4.037497   9.495 1.24e-06 ***
## log(AR)      3.881691   1.225043   3.169  0.00894 **
## DEc         -0.015120   0.004517  -3.348  0.00651 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.071 on 11 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.6804
## F-statistic: 14.84 on 2 and 11 DF,  p-value: 0.0007524
```

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



Residuals vs fitted shows variance moving randomly around the zero value and the normal Q-Q plot shows signs of normality. Coefficients all improved below the 1% significance level.

f) Comment on the coefficient of determination and adjusted coefficeint of determation in the full and final model.

The R^2 value adjusts from 0.7301 in the original to 0.7296 in the new model with two coefficients less. A very minor change of 0.05 for making each coefficient significant at 1%. We would expect minor negative values when changing as the additional variables never decrease the explained variation. The adjusted R^2 of 0.6101 changes to 0.6804. The adjusted R^2 makes use of the ability to compare alternative models compared to the adjusted which makes a bias to models with more explanatory variables. The increase in the adjusted R^2 is more useful as it shows a comparison to the original model without the bias effect the adjusted R^2 has. Can see there is more explained variance in the new model.

g) Compute a 95% Confidence Interval for the AR regression Parameter.

We have df = 14-3=11: $t_{n-3,1-\frac{a}{2}} = t(11, 0.025) = 2.201$

95% CI = $\beta_2 \pm t_{n-3,1-\frac{a}{2}} * s.e.\beta_2$

$= 3.881691 \pm 2.20099 * 1.22504 = (1.185, 6.578)$

We can be 95% confident that the true value for Area of the island will have a long run rate between 1.185 - 6.578, if we drew random smaples repeatedly from the same population and size 95% of the confidence intervals would contain the population value.

Question 2

a) Is the Study balanced or unbalanced.

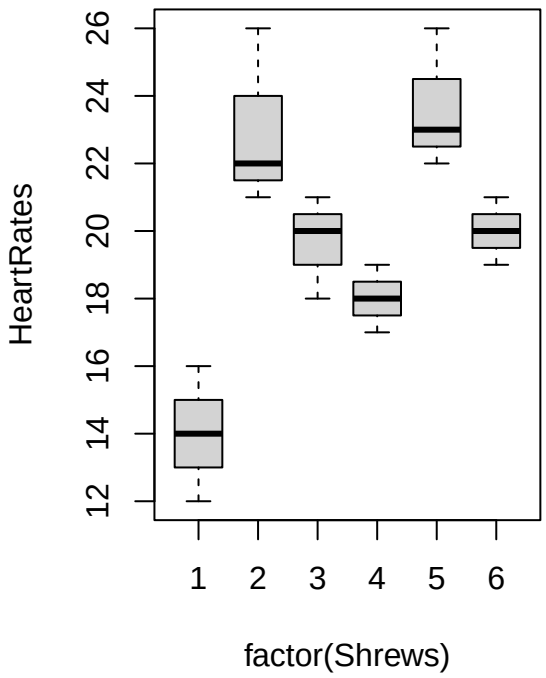
```
TreeShrews <- read.csv("~/R/TreeShrews.dat", sep="")
table(TreeShrews[,c('Shrews','Sleep')])
```

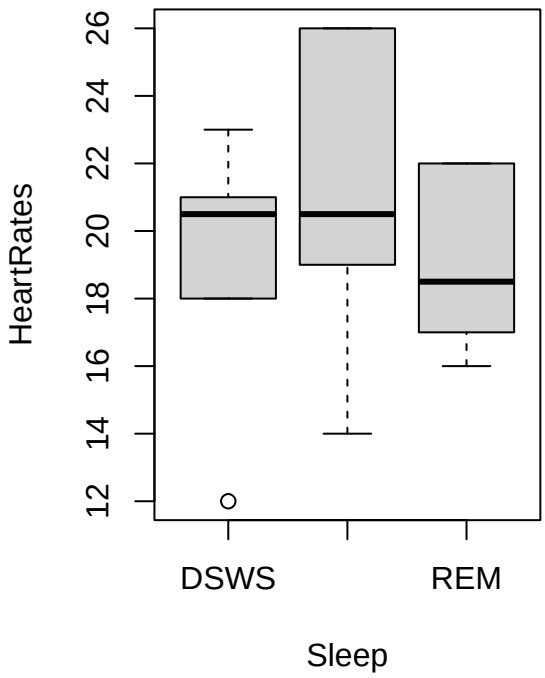
```
##       Sleep
## Shrews DSWS LSWS REM
##      1    1    1   1
##      2    1    1   1
##      3    1    1   1
##      4    1    1   1
##      5    1    1   1
##      6    1    1   1
```

Summing columns we get 6 REM, 6 DSWS and 6 LSWS, data set has no missing observations. Conclude the sample is balanced.

b) Preliminary graphs investigating diferent features.

```
par(mfrow=c(1,2))
boxplot(HeartRates~factor(Shrews),data = TreeShrews);boxplot(HeartRates~Sleep,data = TreeShrews)
```






```
with(TreeShrews, tapply(HeartRates,Shrews,sd))
```

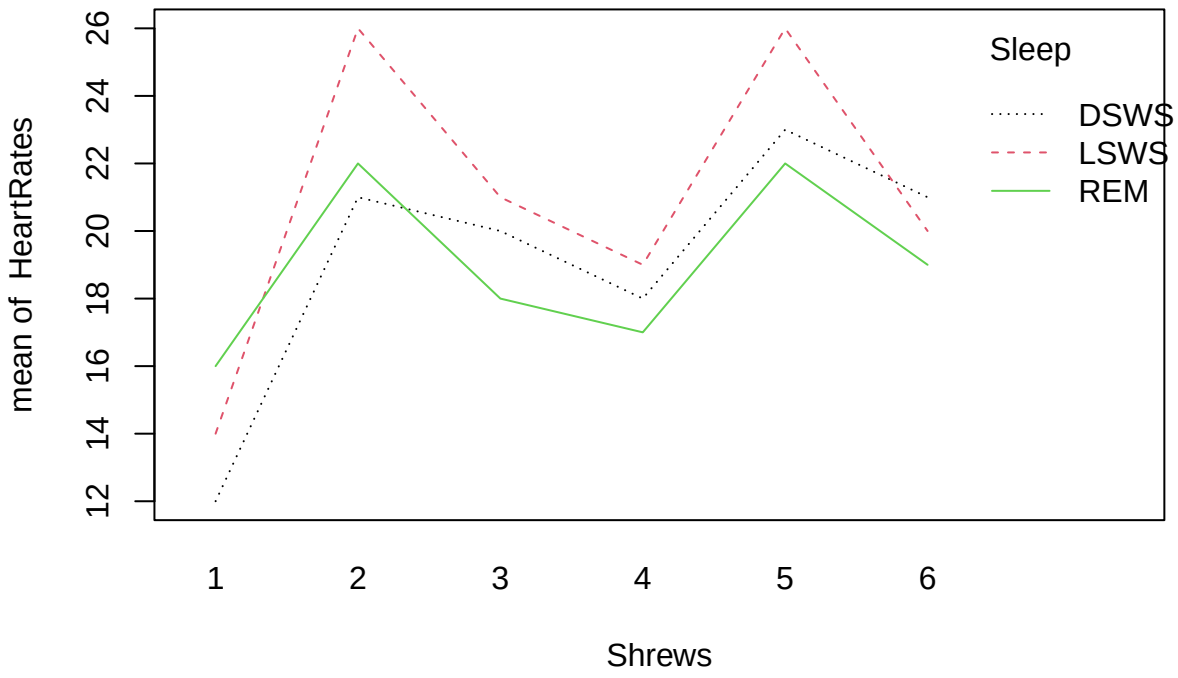
```
##        1        2        3        4        5        6
## 2.000000 2.645751 1.527525 1.000000 2.081666 1.000000
```

```
with(TreeShrews,tapply(HeartRates,Sleep,sd))
```

```
##     DSWS     LSWS      REM
## 3.868678 4.560702 2.529822
```

Shrews appear to have moderately equal variances, only a small number of observations so the sample may be volatile. Can see 1/6 has a large difference in mean heart rate. Standard deviations appear rather similar largest ratio being 2.65/1, slightly over the 2 ratio which is preferential however with the small sample size we can accept equal variances. Sleep variation ratio = 4.56/2.53=1.80 which is also acceptable. Notably DSWS has a large outlier at 12 bpm, adjusting the data set will result in an unbalanced test. Difficult to judge whether with such a small data set whether this observation is an outlier or that more observations should be taken.

```
with(TreeShrews,interaction.plot(Shrews,Sleep,HeartRates,col=1:3))
```



Can see interaction between REM and LSWS, interaction slopes both stepp and thus have minor differences indicating a not so strong interaction. DSWS and LSWS has a large interaction at 6 with large slop differneces. REM and DSWS also appear to intersect at large differing steepness levels.

c) Explain why we cannot fit a two-way ANOVA with interaction model.

```
lm.shrew<-lm(HeartRates~Shrews*factor(Sleep),data=TreeShrews)
anova(lm.shrew)
```

```
## 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
```

$Y = \mu + a_i + \beta_j + \gamma_{ij} + e$

Where Y = mean heart rate, a_i= Shrews, β_j=Sleep, γ_{ij}=interaction

Hypothesis:

$H_0 : \gamma_{ij} = 0, against\ H_1 : \gamma_{ij} \neq 0$

P-value = 0.7271>0.05

Interaction not significant, need to fit reduced model with main effects.

d)Analyse the data

With interaction effect not significant we need to test with the reduced model of the main effects. There is no significant interaction between sleep cycle and Shrews.

Model becomes: $Y = a_i + \beta_j + \epsilon$

```
shrew.2<-lm(HeartRates~factor(Shrews)+Sleep,data=TreeShrews)
anova(shrew.2)
```

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

For main effect of Shrews
$H_O : a_i = 0, H_1 : at\ least\ one\ a_i\ \neq 0$

P-value = 0.0002<0.1

Conclusion: Shrews is significant in explaining Heart Rates.

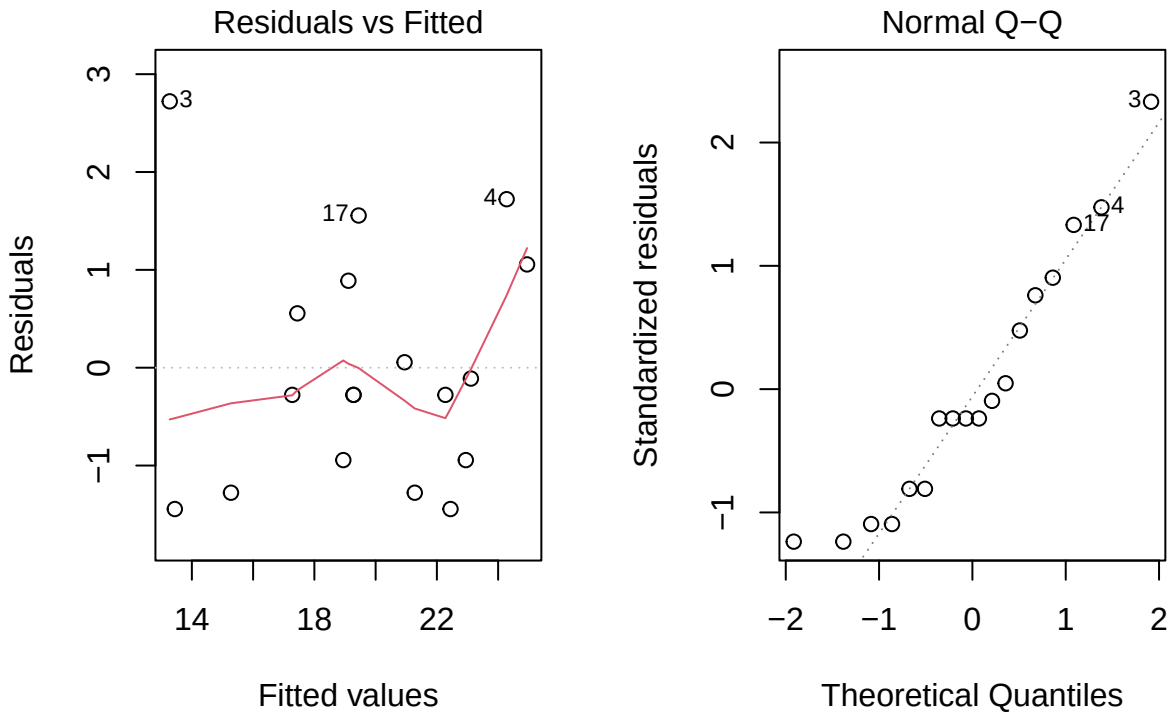
For main effects of Sleep
$H_O : \beta_j = 0, H_1 : at\ least\ one\ \beta_j \neq 0$

P-value = 0.09<.10

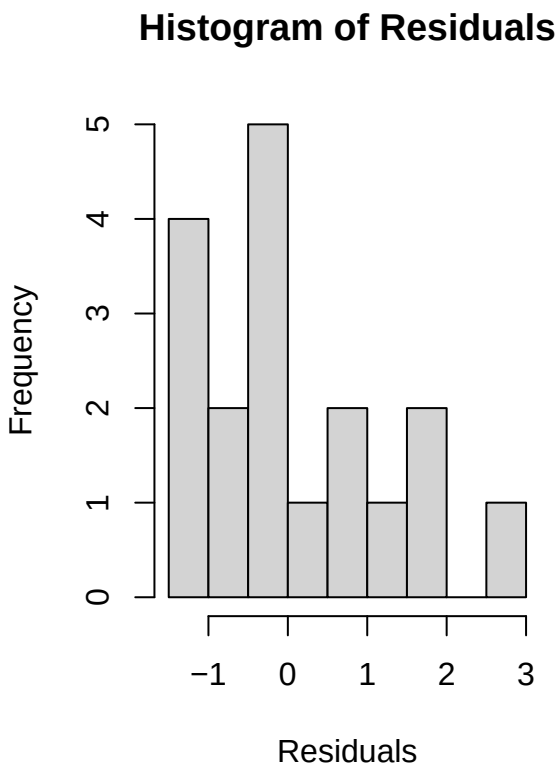
Conclusion Sleep is significant in explaining Heart Rates at the 10% level only.

Validating the model:

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




```
hist(shrew.2$residuals, main = "Histogram of Residuals", xlab = "Residuals")
```

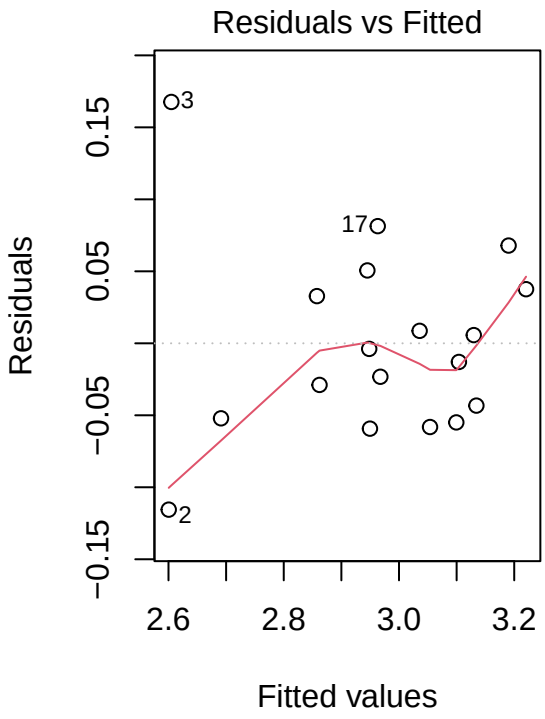


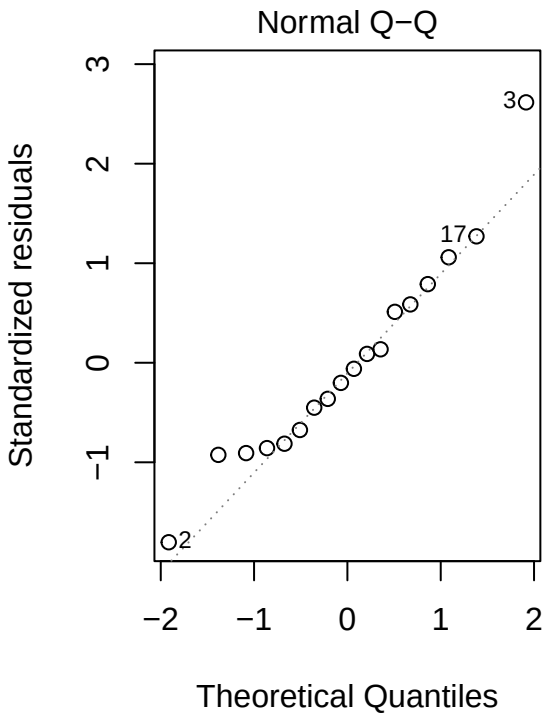
Residuals vs fitted shows a random variance around zero, for the Q-Q plot we can see the influence of the outlier having an impact on the rest of the plot. We would like more observations to make a better decision on normality. Histogram shows a slight right skew of residuals, again with relatively small data set it is hard to judge on the shape of the distribution.The right skew can indicate that a log transform is needed, repeating the test with log of the dependent variable.

```
TreeShrews$lnheart<-log(TreeShrews$HeartRates)
log.lm.shrew<-lm(lnheart~factor(Shrews)+Sleep,data=TreeShrews)
anova(log.lm.shrew)
```

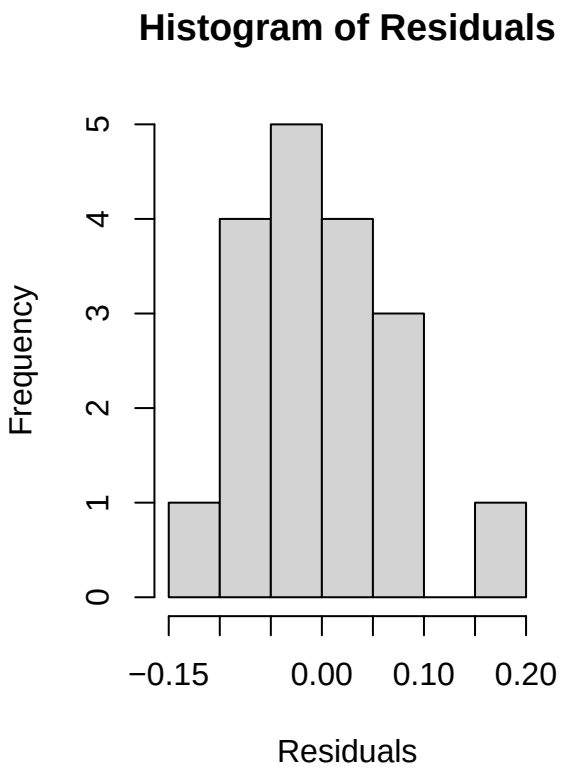
```
## Analysis of Variance Table
##
## Response: lnheart
##                Df  Sum Sq  Mean Sq F value    Pr(>F)
## factor(Shrews)  5 0.55135 0.110269 14.9228 0.0002315 ***
## Sleep           2 0.03137 0.015685  2.1226 0.1704726
## Residuals      10 0.07389 0.007389
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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






```
hist(log.lm.shrew$residuals, main = "Histogram of Residuals", xlab = "Residuals")
```



The log transformed model shows a somewhat randomness of residuals vs fitted and the normal Q-Q plot shows a more linear formation however now impacted by a larger gap of outliers in observation 2 and 3. The histogram is however improved and looks to be approximately normally distributed. Sleep becomes insignificant at the 10% level and the coefficient of determination falls in value.

e) Conclusion

The original model shows significance for both Shrews and Sleep in explaining Heart rates of Shrews. For the most part Shrews appear to display equal variance in heart rates, Sleep appears to have a weaker impact on heart rates than individuals Shrews.A log transformed model improves the histogram of residuals but does not seem to indicate an improved model. What this test seems to convey is that there is a good sign that a large test may prove some interesting results in determining Sleep and Shrews effects on Heart Rate.
