Q1

a.

~~~
```{r}
paramo<- read.table("paramo.dat", header=TRUE)
summary(paramo)
```
~~~

~~~
```{r}
pairs(paramo,panel=panel.smooth)
```
~~~

There is no evident linear relationship between the response variable N and predictors. The relationship between N and DEc is slightly negative and there is mildly positive relationship between N and AR. The rest of predictors almost not have apparent relationship with N.

Among the predictors, there is slightly linear relationship between AR and EL, as well as DEc and DNI.

~~~
```{r}
cor(paramo)
```
~~~

The correlation matrix is the determinant of what we observed in the scatter plot matrix. For instance, the response variable N has no relationship with predictors whereas relationships exist among predictors.

b.

~~~
```{r}
lm1 = lm(N ~ . , data = paramo)
summary(lm1)
```
~~~

$$

N=\beta_1+\beta_2AR+\beta_3EL+\beta_4DEc+\beta_5DNI\\

\hat{Y}-\text{the response(fitted) variable N}\\

X_i:\text{the predictors variables with}\\

X_1=AR;\\

X_2=EL;\\

X_3=DE;\\

X_4=DNI;\\

\epsilon \sin N(0,\sigma^2)\text{denotes the random variation}

$$

The Hypotheses for the Overall ANOVA test of multiple regression

$$

H_0: \beta_3= \beta_4= \beta_5 = 0;\\

H_1: \text{at least one of them do not equal to 0}

$$

```{r}
anova(lm1)
```

F test

```{r}
1 - pf(6.086,4,9)
```

$$

H_0: \beta_2= \beta_3= \beta_4= \beta_5=0\\

H_1: \text{at least one of them do not equal to 0}\\

\text{the Regression SS = 508.92+45.90+537.39+2.06=1094.27}\\

\text{the Mean Square Reg = Reg SS/Reg df = 1094.27/4 = 273.5675}\\

\text{Test statistic}: F:{obs} = MS_{Reg}/MS_{Reg} = 273.5675/44.95=6.086\\

\text{P-value}: P(F_{4,95}\geq 6.086) = 1-pf(6.086,4,9) = 0.008322917 < 0.05;\\

\text{As the P-value less than 0.05, hence we reject }H_0.\\

\text{There is a significant linear relationship between N and at least 4 predictor variables}

$$

~~~
```{r}

summary(lm1)

```
~~~

~~~
```{r}

plot(lm1, which = 1:2)

plot(resid(lm1) ~ AR, data = paramo, ylab = "Residuals"); abline(h=0)

plot(resid(lm1) ~ EL, data = paramo, ylab = "Residuals"); abline(h=0)

plot(resid(lm1) ~ DEc, data = paramo, ylab = "Residuals"); abline(h=0)

plot(resid(lm1) ~ DNI, data = paramo, ylab = "Residuals"); abline(h=0)

```
~~~

From Residuals vs predictor plots, there is no pattern found hence the regression analysis is valid.

From the quantile plot, residuals look linear hence the normal assumption for residuals is appropriate.

From Residuals vs Fitted plot, it demonstrates that the size of the residuals increases as which of fitted values increases, hence the constant variance assumption is inappropriate.

d.

~~~
```{r}

summary(lm1)$r.squared

```
~~~

$$
R^2 = \frac{S.S._{Regression}}{S.S._{Total}} =
\frac{S.S._{Total} - S.S._{Residuals}}{S.S._{Total}} = 0.730068
$$

R-square means the goodness of fit measure for linear regression models. It represents the proportion of the variance for the dependent variable which is explained by an independent variable or variables in a regression model.

e.

```{r}
step(lm1, direction = "backward")
```

$$
\text{The most appropriate model is: N = 30.79797 + 6.68304AR - 0.01706DEc}
$$

```{r}
lm2 = lm(formula = N ~ AR + DEc, data = paramo)
summary(lm2)
```

f.

```{r}
summary(lm2)$r.squared
```

R-square remains the same or increases if there are additional input variables. In the regression 1, both R-squares are nearly the same after dropping 2 variables hence the input variable of temperature is not involved to explain the output variable. The adjusted model is more fit.

g.

```{r}
confint(lm2, "AR", level = 0.95)
```

We are 95% confident that the regression parameter of AR coefficients is between 1.699121 and 11.66696.

Q2

a.

```{r}
tree = read.table("TreeShrews.dat", header = TRUE)
head(tree)
```

```{r}
table(tree[, c("Sleep", "Shrews")])
```

The numbers of replicates are equal hence the design is balanced.

b.

```{r}
with(tree, interaction.plot(Shrews, Sleep, HeartRates, col = 1:3, fixed = TRUE))
```

c.

~~~
$$

Y = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon\\

Hypothesis: H_0: \gamma{ij} = 0; H_1: \text{at least 1}\gamma_{ij}\text{is not 0}

$$

```{r}

anova(lm(HeartRates ~ Shrews * Sleep, data = tree))

```


$$

\text{from anova table}\\

P-value = 0.7271 > 0.05\\

\text{the interaction is insignificant, so we will fit reduced model with main effects only}

$$

d.

$$

\text{General model}: Y = \mu + \alpha_i + \beta_j +\epsilon \sim^{i.i.d}

N(0, \sigma^2)

Y = \text(Heart rates response)\\

\mu:\text{Overall mean response}\\

\alpha_i:\text{Effect difference in different shrews}\\

\beta_j:\text{effect of Sleep types}

\epsilon:\text{Unexplained variation}

$$

```{r}

tree.1 = lm(HeartRates ~ Shrews * Sleep, data = tree)

tree.2 = update(tree.1, . ~ . -Shrews:Sleep)

```


```{r}
~~~

anova(tree.2)

$$
\text{Main effect:Sleep}\\
Y = \mu + \alpha_i +\beta_j + \epsilon\\
\text{Hypothesis}: H_0:\beta_j = 0; H_1:\text{at least 1}\beta_j \text{is not 0}\\
P-value = 0.56051 > 0.05\\
\text{Sleep type is insignificant}
$$

$$
\text{Main effect:Shrews}\\
Y = \mu + \alpha_i +\alpha_i + \epsilon\\
\text{Hypothesis}: H_0:\alpha_i = 0; H_1:\text{at least 1}\alpha_i \text{is not 0}\\
P-value = 0.09432 > 0.05\\
\text{There is insignificant effect on different shrews}
$$

```{r}
par(mfrow = c(1,2))
plot(tree.2, which = 1:2)
```

There is mild curvature in the normal quantile plot of residuals. From the residual plot, we observe that smaller variability near the fitted values at 23. In conclusion, residuals look closer to normally distributed.

```{r}
tree.main = lm(log(HeartRates) ~ Shrews + Sleep, data = tree)
anova(tree.main)
```

e.

Both shrews and sleep effects are insignificant when they fit the log of heart rates. There are no obvious effects when lines are nearly parallel from the plot in b. From the hypothesis test of d, the P-value demonstrates that the effect is insignificant.