Stat 680 Assignment 1

Sukhdeep Singh

May 27, 2018

Solution 1: Reading the file

```
paramo=read.table('paramo.dat', header = T)
```

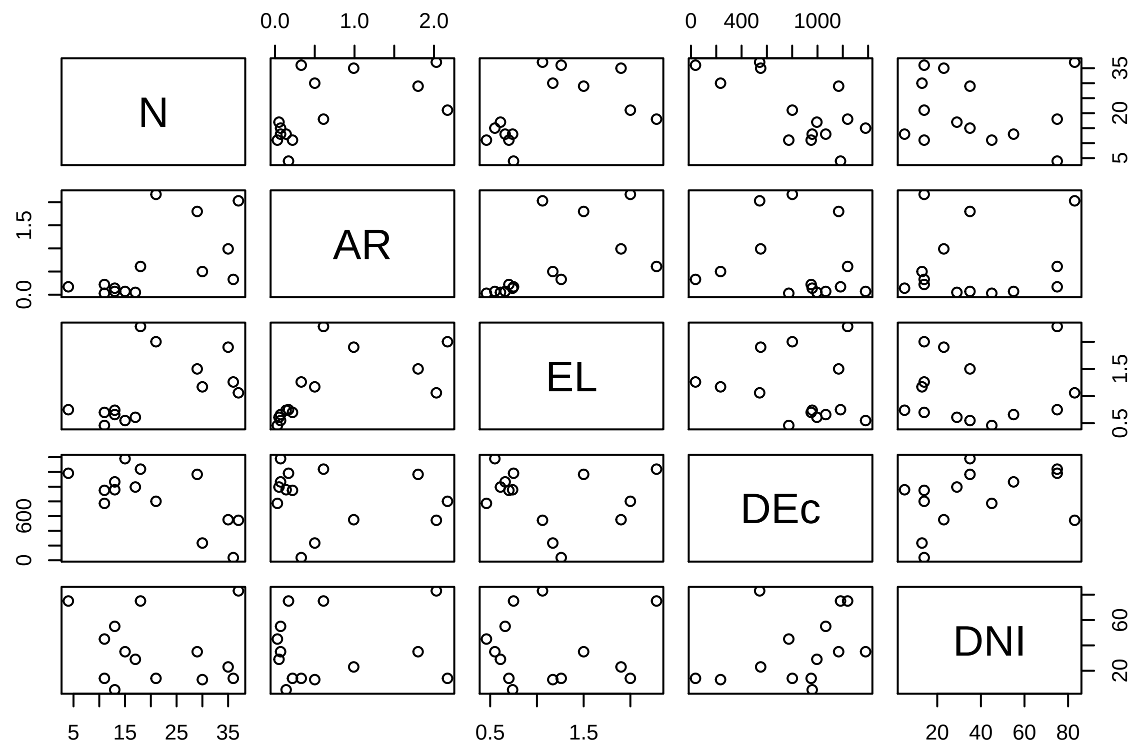
Reponse variable: N (number of birds)

List of predictors:

- AR (area of island)
- EL (Elevation)
- DEc (Distance from Ecuador)
- DNI (Distance to the nearest island)

A) Plotting the scatterplot to study the relationship between the predictors and response variables:

```
plot(paramo)
```



Looking at correlation matrix

```
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
```

The response variable (N) shows a moderate postive correlation with AR and relatively stronger negative correlation with DEc. Correlation with other predictors is noticebaly weak.

We can see a fair bit of positive correlation between AR and EL, rest of the interactions between predictors are weak.

B) Creating a fully fitted linear model with all the predictors

The mathematical multiple regression model: $N = B_0 + B_1(AR) + B_3(EL) + B_3(DEc) + B_4(DNI) + E\mu$

- B_0:Intercept of regression equation
- B_1,B_2,B_3,B_4 : coefficients of the predictors
- Eµ : unexplained random variation

```r
full_model=lm(N ~ AR + EL + DEc + DNI, data = paramo)

summary(full_model)
```

```
##
## 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
```

Hypothesis for anova

- H_0: B_1=B_2=B_3=B_4=0
- H_1: Alteast one of the coeffients is non zero

Anova table

```r
aov=anova(full_model)

print(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
```

F statistic and P-value

F-statistic: 6.085 on 4 and 9 DF, p-value: 0.01182

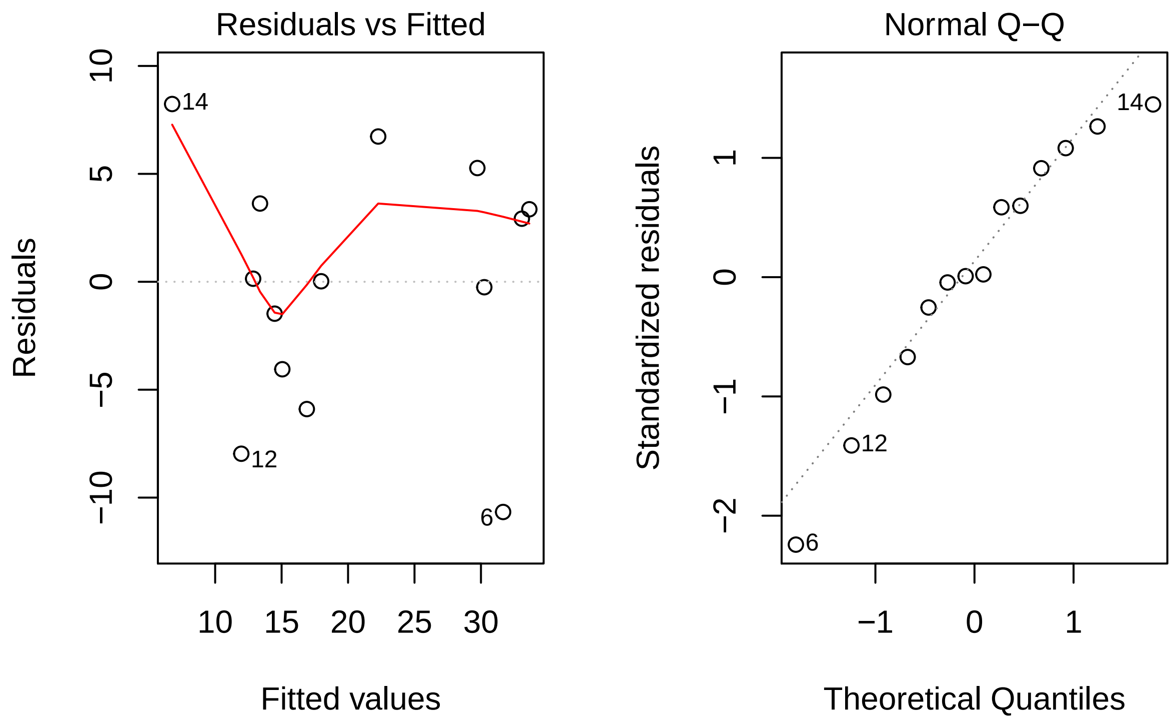
Null distribution

F(4,9)

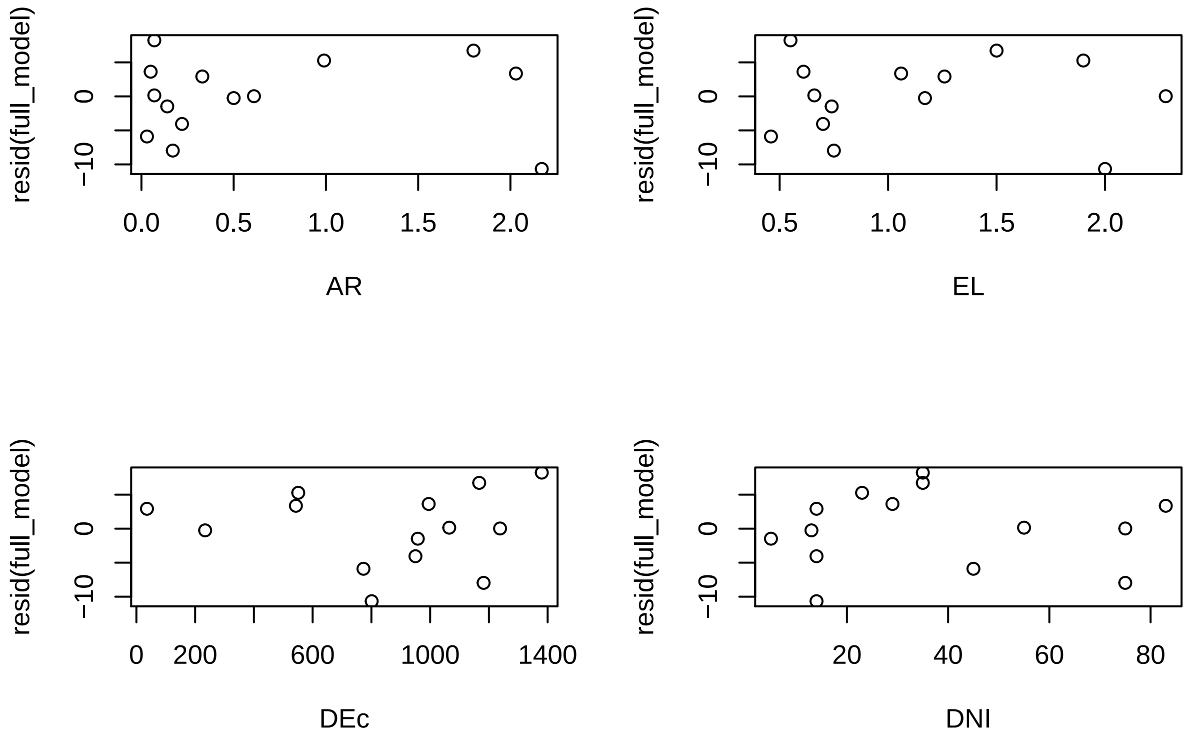
Therefore we can conclude that there exists a significant relationship betweent the response and predictors.

C) Validating assumptions

```r
par(mfrow = c(1, 2))
plot(full_model, which=1:2)
```




```r
par(mfrow = c(2, 2))
plot(resid(full_model)~AR+EL+DEc+DNI, data =paramo)
```




```r
summary(full_model)
```

```
##
## 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
```

Conclusion: QQ-norm plot of redisuals is linear .This confirms the assumption normal distributed residuals holds. Also the residuals are equally distributed for the predictors as seen above. Summary of the model indicates inclusion of many insignificant predictors which can be pruned further to improve the prediction.

D) R squared: Multiple R-squared: 0.7301, , Adjusted R-squared: 0.6101

This term indicates the goodness of fit of the model, how well the regression sum of squares explain the variation. Adjusted R sqaure, penalises on the basis of number of parameters fitted in the model.

E) Removing the predictor with least significance and fitting the model again

```
three_model=lm(N ~ AR + EL + DEc, data = paramo)

summary(three_model)
```

```
##
## 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
```

```
anova(three_model)
```

```
## Analysis of Variance Table
##
## Response: N
##           Df Sum Sq Mean Sq F value   Pr(>F)
## AR         1 508.92  508.92 12.5151 0.005378 **
## EL         1  45.90   45.90  1.1288 0.313020
## DEc        1 537.39  537.39 13.2152 0.004572 **
## Residuals 10 406.65   40.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From summary of the regression model we can see some improvement in the adjusted R-squared to 0.6473, and new anova F-test confirms model is still significant p-Value of 0.003499. We can further improve the model by removing insignificant predictors

```
second_model=lm(N ~ AR + DEc, data = paramo)

summary(second_model)
```

```
##
## 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
```

```
anova(second_model)
```

```
## 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
```

From the summary of the regression model we can see that adjusted R-square has increased further to 0.6588 and anova table indicates that both predcitors in the model are significant. Hence we can conclude that we have arrived at parsimonous model, where response variable (N) can be best explained by AR and DEc.

F) R-squared for final model: 0.7113 Adjusted R-squared for final model: 6588 Upon increasing the number of predictors in the regression model, the R-squared value increases, however, adjusted R-square takes into account the number of predictors in the model and penalises the R-sqaured value based on the number of predictors.

G)

```
x=qt(0.975,11)
upperbound=6.683038+x*2.264403
lowerbound=6.683038-x*2.264403

print(lowerbound)
```

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

```
print(upperbound)
```

```
## [1] 11.66696
```

95% Confidence interval for AR is (0.1879308, 13.17815)

Solution 2: Reading the data

```
battery=read.table('powercell.dat', header = T)
```

A)Mathematical polynomial model to fit: B_0+B_1(charge) + B_2(Temp) + B_{11}(charge^2) + B_{22}(temp^2) + B_{12}(charge*temp)+ E (unexplained variation)

```r
big_model=lm(cycle ~ charge + temp + I(charge^2) + I(temp^2) + I(charge*temp), data = battery)

summary(big_model)
```

```
##
## Call:
## lm(formula = cycle ~ charge + temp + I(charge^2) + I(temp^2) +
##     I(charge * temp), data = battery)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.40 -110.97   15.72  120.83  366.18
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      380.27122  175.43209   2.168 0.033123 *
## charge          -763.80347  308.93712  -2.472 0.015516 *
## temp              13.66182    3.63606   3.757 0.000323 ***
## I(charge^2)      117.20961  124.57577   0.941 0.349569
## I(temp^2)         -0.20825    0.05572  -3.737 0.000345 ***
## I(charge * temp)   4.78291    2.86490   1.669 0.098882 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 185.8 on 81 degrees of freedom
## Multiple R-squared:  0.6422, Adjusted R-squared:  0.6201
## F-statistic: 29.07 on 5 and 81 DF,  p-value: < 2.2e-16
```

Equation for model 1: 380.27 - 763.80(Charge) + 13.66(Temp) + 117.209(charge^2) -0.20825(Temp^2) + 4.78291(charge * temp)

Linear model: $B_0 + B_1$(charge) + B_2(Temp) + E (unexplained variation)

```r
linear_Model=lm(cycle ~ charge +temp, data = battery)

summary(linear_Model)
```

```
##
## Call:
## lm(formula = cycle ~ charge + temp, data = battery)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -457.95 -111.21   42.05  132.02  355.34
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  153.826     63.178   2.435    0.017 *
## charge      -409.695     51.486  -7.957 7.38e-12 ***
## temp          12.493      1.306   9.563 4.37e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 198.2 on 84 degrees of freedom
## Multiple R-squared:  0.5779, Adjusted R-squared:  0.5678
```

```
## F-statistic:  57.5 on 2 and 84 DF,  p-value: < 2.2e-16
```

Linear equation: 153.826 - 409.695(charge) + 12.493(temp)

B) Covariance of the data:

```
cov(battery)
```

```
##               cycle        temp      charge
## cycle  90892.38690 2692.287298 -46.1101258
## temp    2692.28730  307.056402   2.7919754
## charge   -46.11013    2.791975   0.1976863
```

Regression SS = (12.483)(87-1)(2692.2872) + (- 409.695)(87-1)(-46.1101258) = 4514906.18

C) Comparison of Multiple linear model against the polynomial model

H_0: Residual SS are same for both models (Same explanatory power)

H_1: Residual SS are not same for both the model

```
anova(big_model,linear_Model)
```

```
## Analysis of Variance Table
##
## Model 1: cycle ~ charge + temp + I(charge^2) + I(temp^2) + I(charge *
##     temp)
## Model 2: cycle ~ charge + temp
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     81 2797003
## 2     84 3299456 -3   -502453 4.8503 0.003735 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova result we observe a p-value of 0.00375 which means we will be rejecting null hypothesis and this means, polynomial model has higher explanotory power of the variation as compared to the linear model.

Solution 3 Reading the data:

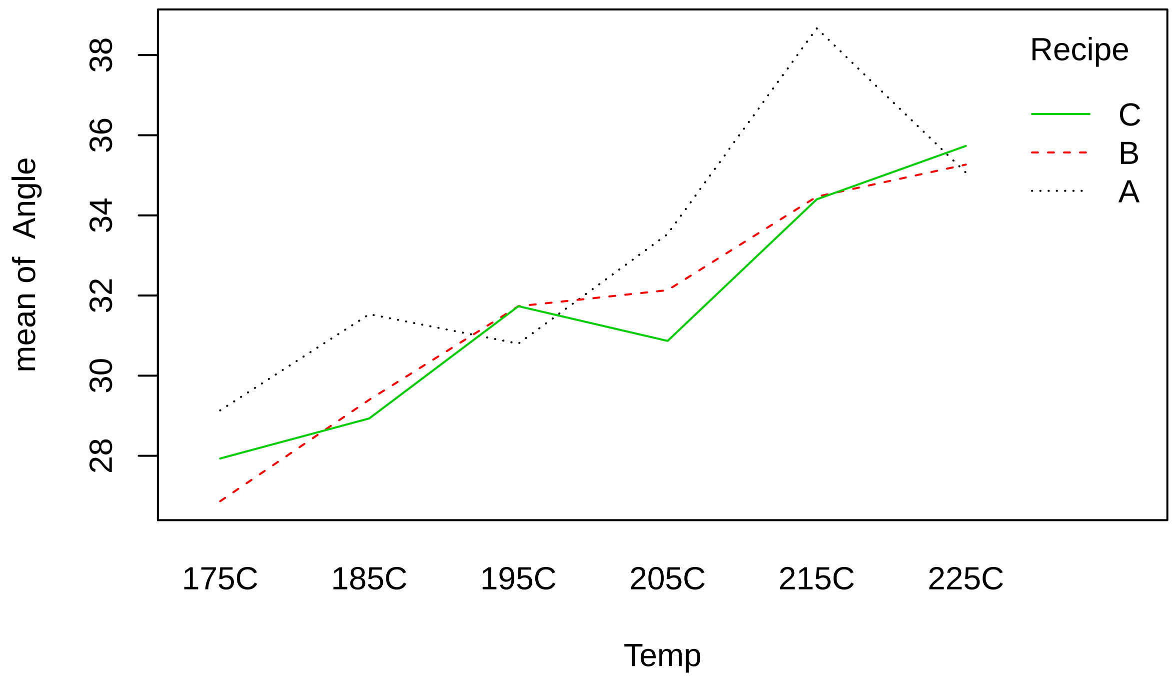
```
cakedata=read.table('cake.dat', header = T)

table(cakedata[,1:2])
```

```
##       Recipe
## Temp    A  B  C
##   175C 15 15 15
##   185C 15 15 15
##   195C 15 15 15
##   205C 15 15 15
##   215C 15 15 15
##   225C 15 15 15
```

This is an example of balanced study as the number of replicates across all the factors are equal.

b) Preliminary plots for chekcking the interaction

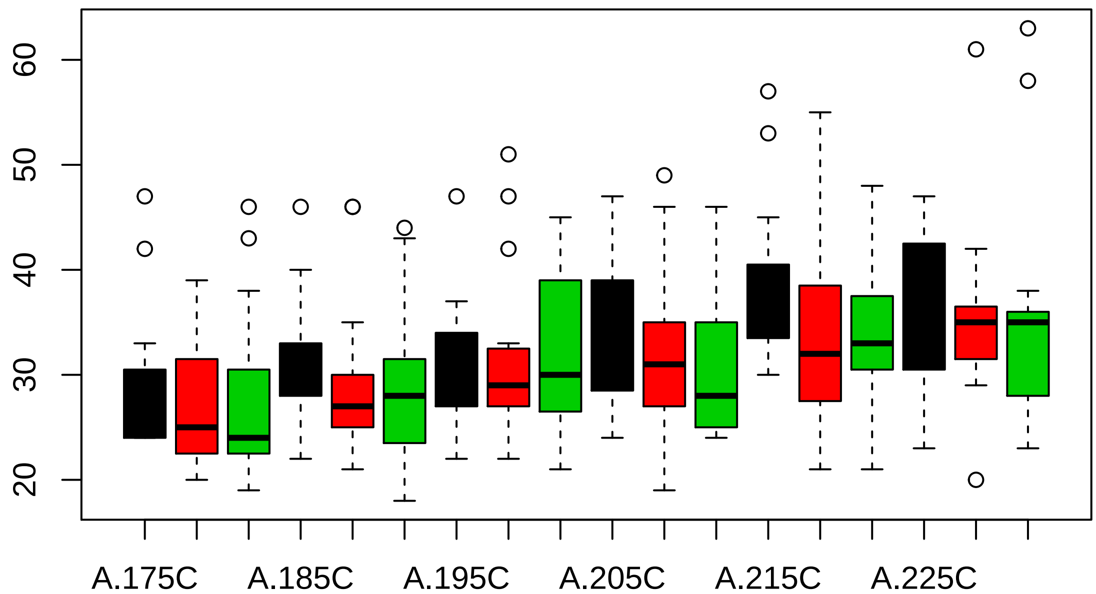
```
with(cakedata,interaction.plot(Temp,Recipe,Angle, col = 1:3))
```



From the above interaction plot we can confirm that there is a some interaction between the temperature at 195 and recipes A,B and C. We see that interaction of recipe C and A becomes weak between tempratures (175-185 and 195-215)

Next we check the boxplot

```
boxplot(Angle ~ Recipe + Temp, data = cakedata,col=1:3)
```



The above boxplot depicts the variation of response variable (Angle) explained by factors Recipe + Temp

- Recipe A: Denoted in black
- Recipe B: Denoted in Red
- Recipe C: Denoted in green

We can see some outliers in the data throughout the observations, for all recipes and temperatures. The variability of breaking angle is fairly similar for the 3 recipes, with recpie B showing highest variation at temp 215

c) Model can be given by:

Y_{angle}= B_0 + B_{recipe} + B_{temp} + B~recipe*temp~ + E(unexplained variation)

Where B_{recipe} is the effect due to the factor (recipe) B_{temp} is the effect due to the factor (temp) B~recipe*temp~ is the interaction between factors (recipe & temp) E is the unexplained variation

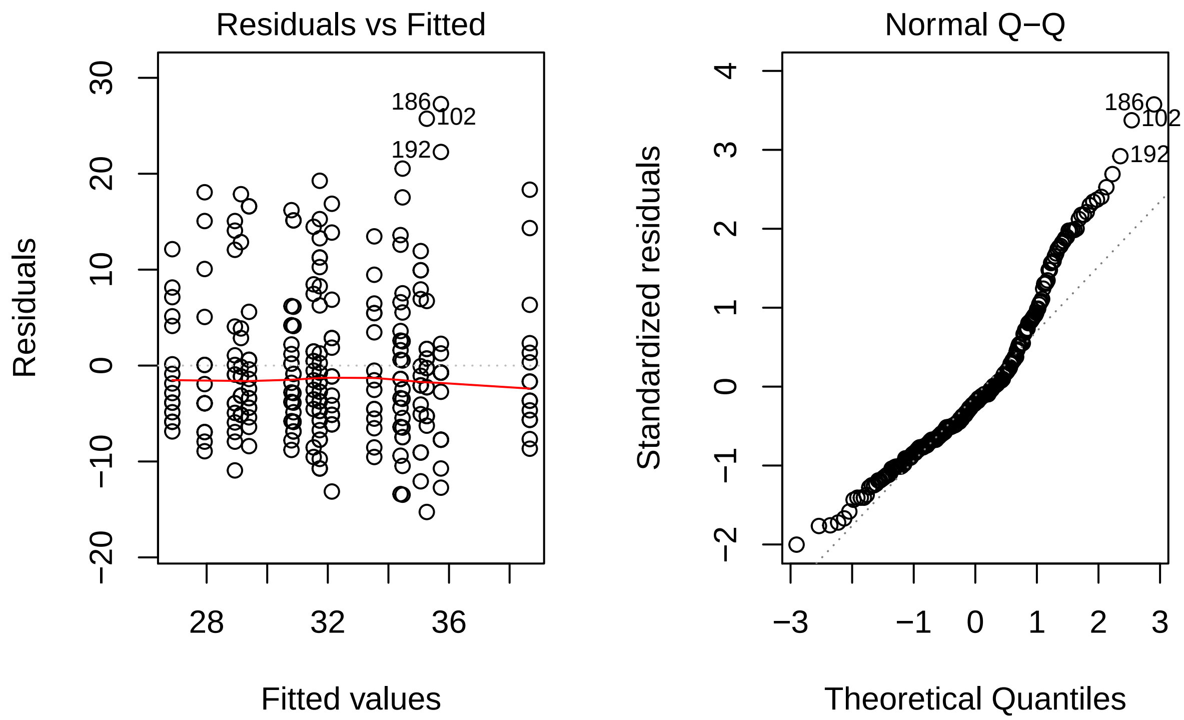
Hypothesis for the model

H_0: B(recipe*temp) is zero (no effect of predictor interaction on response)

H_1: B(recipe*temp) is not zero

Validating the assumptions

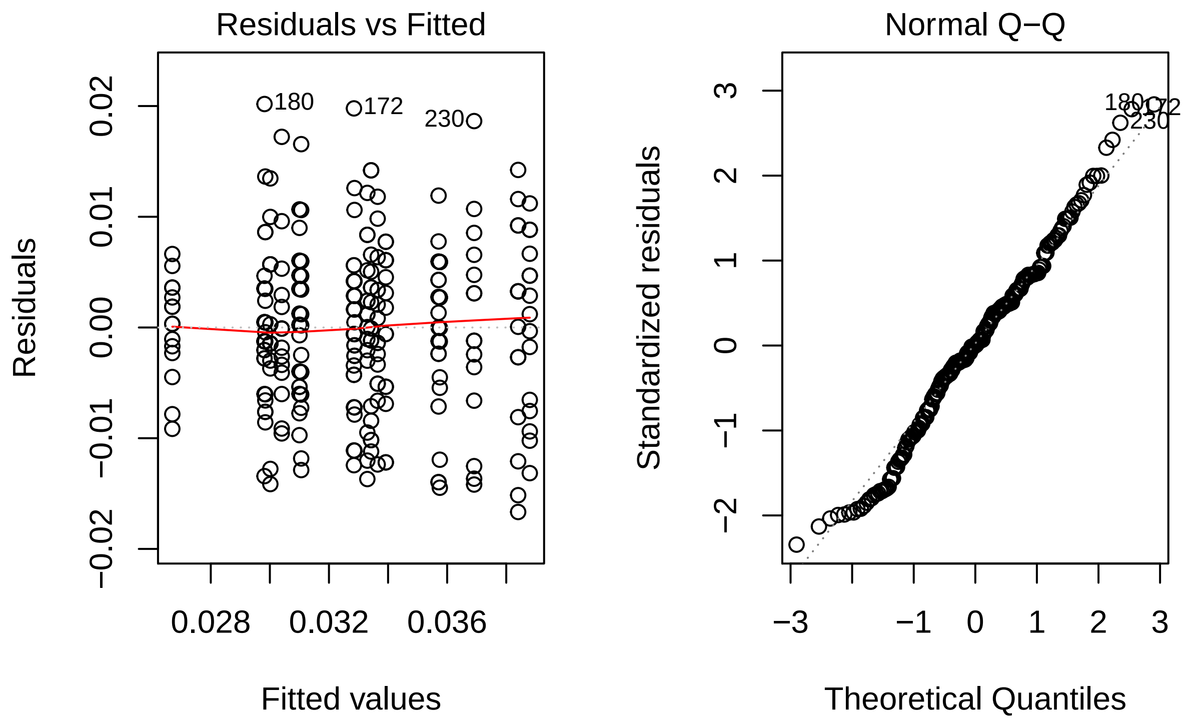
```
cakemodel= lm(cakedata$Angle ~  factor(cakedata$Recipe)*factor(cakedata$Temp))
par(mfrow=c(1,2))
plot(cakemodel, which = 1:2)
```



We can see the fitted values are below the residuals=0 line and the QQ-plot shows some curvature. Hence we will attempt to apply a tranformation to the response variable.

Applying the inverse tranformation on the response variable, we can see that the residuals are much more evenly distributed than before and QQ plot shows a fairly linear nature.

```
cakemodel = lm(Angle^(-1) ~ factor(cakedata$Recipe)*factor(cakedata$Temp), data = cakedata)
par(mfrow=c(1,2))
plot(cakemodel, which = 1:2)
```



Conducting the anova test on the transformed model

```
anova(cakemodel)
```

```
## Analysis of Variance Table
##
## Response: Angle^(-1)
##                                              Df    Sum Sq    Mean Sq
## factor(cakedata$Recipe)                       2 0.0002634 0.00013168
## factor(cakedata$Temp)                         5 0.0022220 0.00044441
## factor(cakedata$Recipe):factor(cakedata$Temp)  10 0.0001813 0.00001813
## Residuals                                   252 0.0136515 0.00005417
##                                             F value    Pr(>F)
## factor(cakedata$Recipe)                      2.4307   0.09004 .
## factor(cakedata$Temp)                        8.2036 3.399e-07 ***
## factor(cakedata$Recipe):factor(cakedata$Temp)  0.3347   0.97109
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see a p-Value of 0.97 for the interaction coeffiecient hence we cannot reject the null hypothesis.

d) We can see from interaction plots interaction among the predictors is weak in nature and variability of response variable is more or less simlar across the factors with some exceptions, anova test of model also confirms that temperature is the only significant predictor for the response variable.