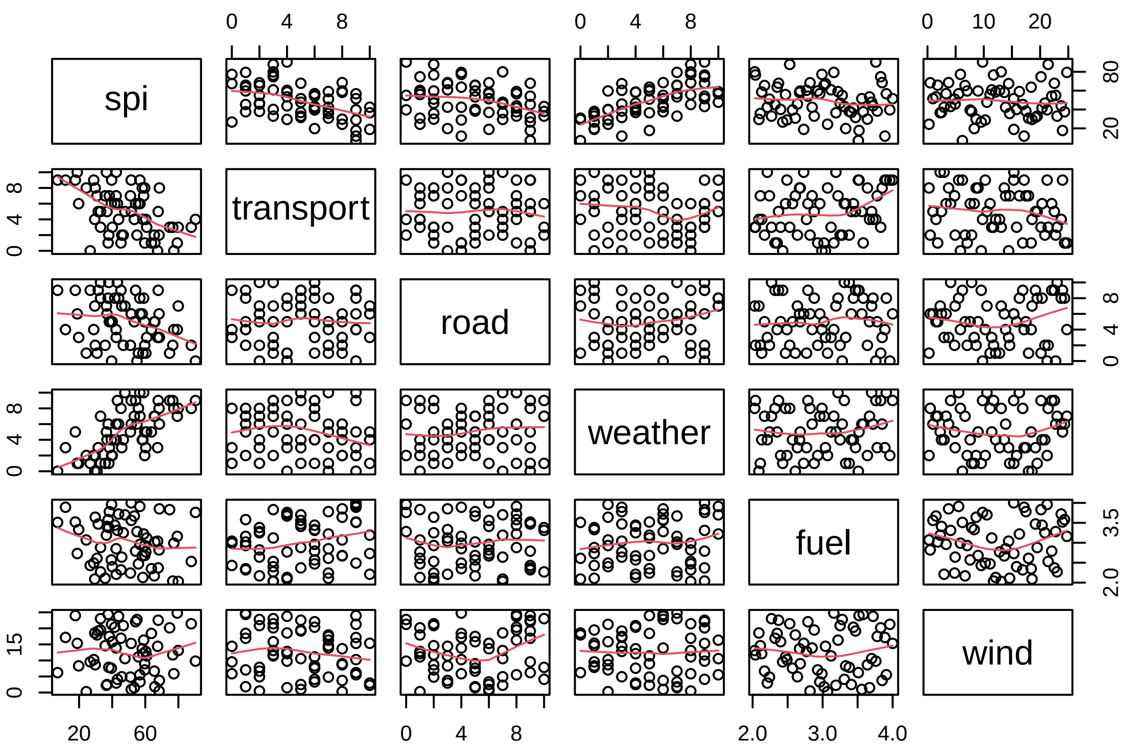
# RStudio Project Report

Author: The Vinh (Vin) Ha

**Study on Traffic dataset:**

**a) Correlation plot and matrix, comments on relationships of predictors and response:**

```
traffic = read.csv('data/traffic.csv', header = TRUE)
pairs(traffic, panel = panel.smooth)
```




```
cor(traffic)
```

```
##                  spi    transport         road     weather         fuel
## spi        1.00000000 -0.472909967 -0.303836850  0.66672345 -0.138153417
## transport -0.47290997  1.000000000 -0.005714728 -0.16971072  0.240947972
## road      -0.30383685 -0.005714728  1.000000000  0.12495993  0.043675635
## weather    0.66672345 -0.169710717  0.124959926  1.00000000  0.110531767
## fuel      -0.13815342  0.240947972  0.043675635  0.11053177  1.000000000
```

```
## wind      -0.03466263 -0.131014749  0.080481857  0.00751783  0.006532832
##                   wind
## spi       -0.034662632
## transport -0.131014749
## road       0.080481857
## weather    0.007517830
## fuel       0.006532832
## wind       1.000000000
```

- The response variable ***spi*** has a strong negative relationship with the predictor ***transport***; a weak negative relationship with the predictor ***road***; a strong positive relationship with the predictor ***weather***; and no obvious relationship with both the predictors ***fuel*** and ***wind***.
- There does not seem to be a relationship between the predictors themselves.

### b) Fit full model and estimate the impact of *weather* on *spi* with 95% CI:

```r
M1 = lm(spi ~ ., data = traffic)
summary(M1)
```

```
##
## Call:
## lm(formula = spi ~ ., data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1596  -4.9415   0.1278   5.1686  21.7415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.8071     7.4080   8.478 1.27e-11 ***
## transport    -2.1750     0.4611  -4.717 1.63e-05 ***
## road         -2.4097     0.4365  -5.520 9.04e-07 ***
## weather       4.2456     0.4473   9.492 2.92e-13 ***
## fuel         -3.6145     2.2759  -1.588    0.118
## wind         -0.1358     0.1764  -0.769    0.445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.913 on 56 degrees of freedom
## Multiple R-squared:  0.7405, Adjusted R-squared:  0.7174
## F-statistic: 31.96 on 5 and 56 DF,  p-value: 3.039e-15
```

```r
summary.M1 = summary(M1)
sqrt(diag(summary.M1$cov.unscaled * summary.M1$sigma^2))[4]
```

```
##   weather
## 0.4472731
```

```r
qt(1 - 0.05/2,56)
```

```
## [1] 2.003241
```

**The require CI is:**

$$\hat{\beta}_{\text{weather}} \pm t_{n-p,1-\frac{\alpha}{2}} \text{se}(\hat{\beta}_{\text{weather}}) = \hat{\beta}_{\text{weather}} \pm t_{56,0.975} \text{se}(\hat{\beta}_{\text{weather}}) = 4.2456 \pm 2.003241 \times 0.4472731 = (3.349604, 5.141596)$$

That is, we are 95% confident that for every percentage increase in relative ***weather***, the ***spi*** concentration will increase between **3.349604** and **5.141596** on average.

**c) Conduct F-test for overall regression and examine relationship between predictors and response:**

**Theorical Model is:**

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 Xi5 + \epsilon_i, \quad i = 1, 2, \ldots, n$$

- $Y_i$ is the response variable ***spi***
- $X_{ij}$ is th are the predictors variables for the i-th observation:
  - $X_{i1}$ = annual mean ***transport*** of test locations
  - $X_{i2}$ = annual mean ***road*** of test locations
  - $X_{i3}$ = annual mean ***weather*** of test locations
  - $X_{i4}$ = annual mean ***fuel*** of test locations
  - $X_{i5}$ = annual mean ***wind*** of test locations
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ denotes the random variation with constant variance

**Now we conduct the F-test:**

- Hypotheses: $H_0 : \beta_1 = \ldots = \beta_5 = 0$ vs $H_1$ : not all $\beta_i = 0$, for $i = 1, 2, \ldots, 5$
- Standard R output ANOVA table:

```
anova(M1)
```

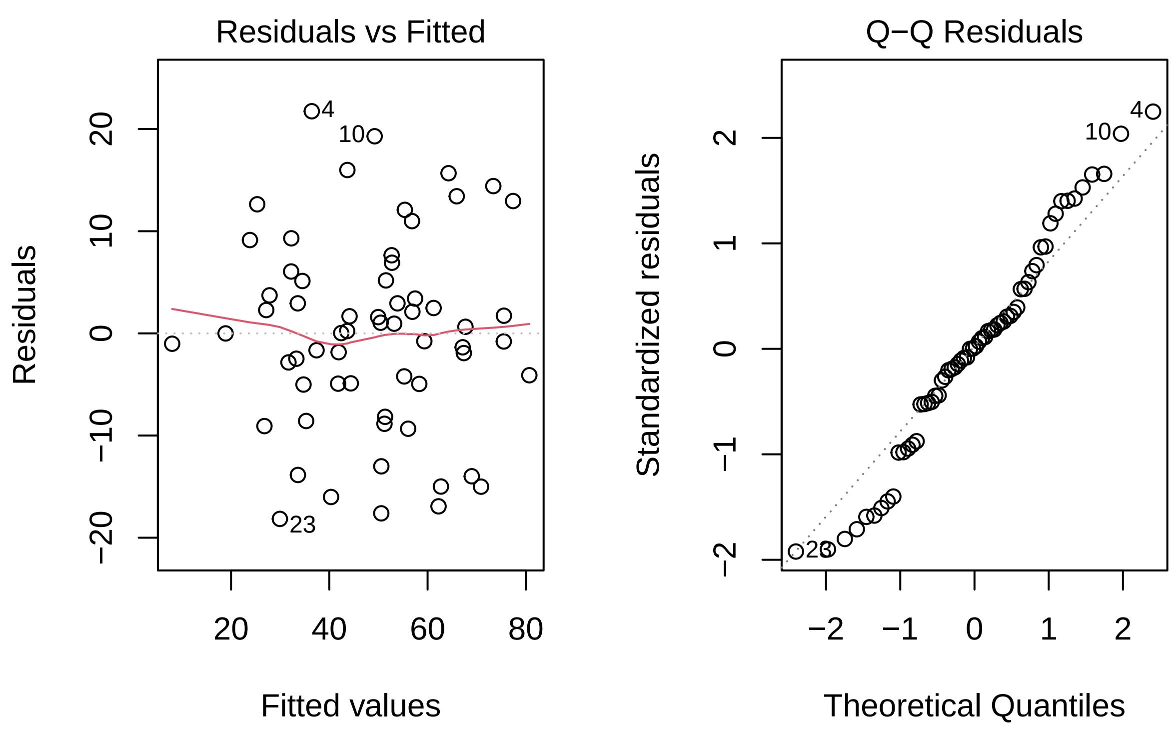
```
## Analysis of Variance Table
##
## Response: spi
##           Df Sum Sq Mean Sq F value    Pr(>F)
## transport  1 4742.6  4742.6 48.2656 4.228e-09 ***
## road       1 1992.7  1992.7 20.2800 3.441e-05 ***
## weather    1 8651.9  8651.9 88.0507 4.355e-13 ***
## fuel       1  258.1   258.1  2.6264    0.1107
## wind       1   58.2    58.2  0.5921    0.4449
## Residuals 56 5502.6    98.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Note the $RegressionSS = 4742.6 + 1992.7 + 8651.9 + 258.1 + 58.2 = 15703.5$
- Therefore the Mean Squared Reg = Reg SS / Reg df = 15703.5/5 = 3140.7
- Test statistics: $F_{obs} = MS_{Reg}/MS_{Res} = 3140.7/98.26094 = 31.96285$
- The null distribution for the test statistics is: $F_{5,56}$
- P-value: $P(F_{5,56} \geq 31.96285) = 0 = 3.0386681751212669e - 15 < 0.05$
- As the P-value is small:

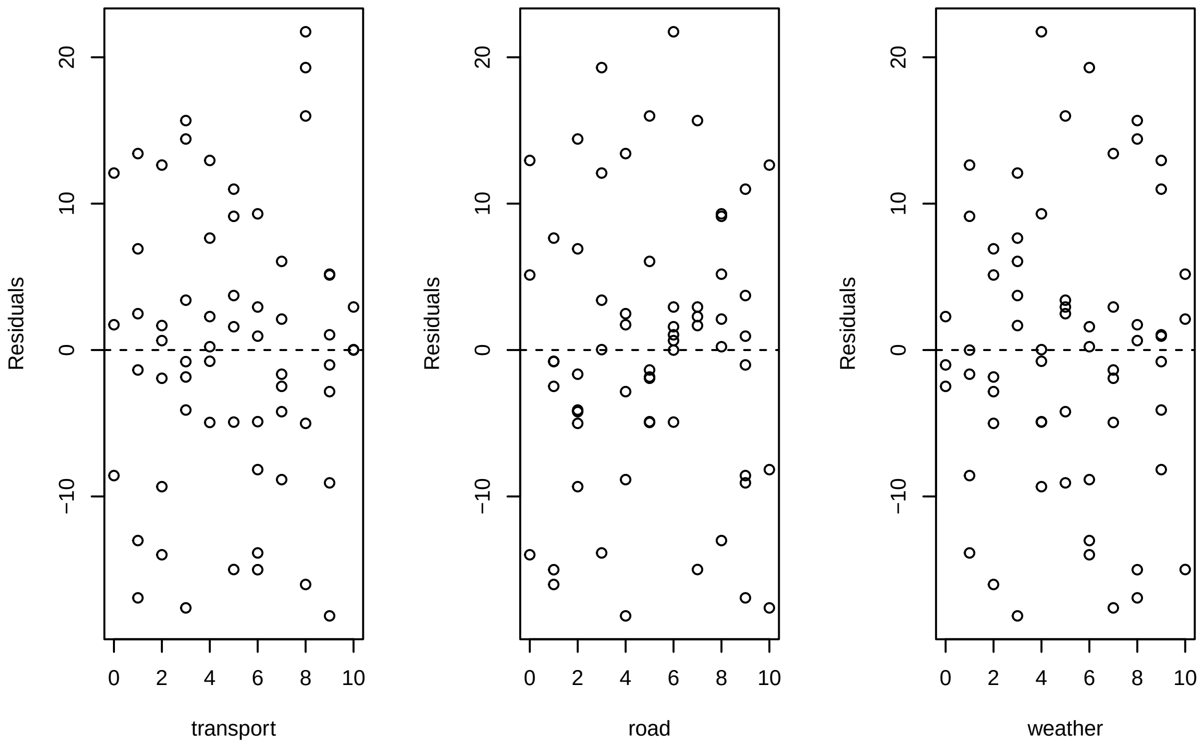
- (Statistical) There is enough evidence to reject $H_0$
- (Contextual) There is significant linear relationship between ***spi*** and at least one of the 4 predictor variables.

**d) For the diagnostics:**

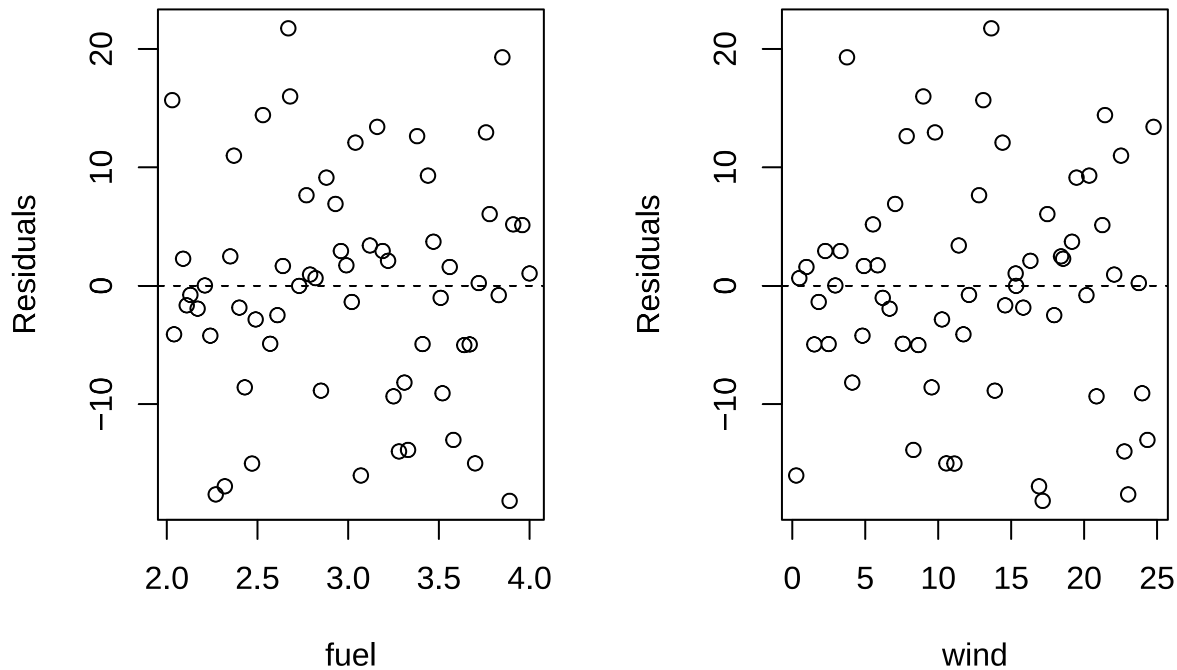
```
par(mfrow = c(1,2))
plot(M1, which = 1:2)
```




```
par(mfrow = c(1,3))
plot(resid(M1) ~ transport, data = traffic, xlab = "transport", ylab = "Residuals")
abline(h = 0, lty = 2)
plot(resid(M1) ~ road, data = traffic, xlab = "road", ylab = "Residuals")
abline(h = 0, lty = 2)
plot(resid(M1) ~ weather, data = traffic, xlab = "weather", ylab = "Residuals")
abline(h = 0, lty = 2)
```




```
par(mfrow = c(1,2))
plot(resid(M1) ~ fuel, data = traffic, xlab = "fuel", ylab = "Residuals")
abline(h = 0, lty = 2)
plot(resid(M1) ~ wind, data = traffic, xlab = "wind", ylab = "Residuals")
abline(h = 0, lty = 2)
```




- The quantile plot of residuals look approximately linear, so the normality assumption for residuals is appropriate
- There is no obvious pattern in any of the residual plots so it appears the linearity and constant variance assumptions of the multiple linear model are

**e) Find R2:**

- Here $R^2 = 0.741 = 74.1\%$, which is a goodness of fit metric. It means 74,1% of the variation in ***spi*** is explained by the full linear regression model.

**f) Find best regression model:**

```r
summary(M1)
```

```
##
## Call:
## lm(formula = spi ~ ., data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1596  -4.9415   0.1278   5.1686  21.7415
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.8071     7.4080   8.478 1.27e-11 ***
## transport    -2.1750     0.4611  -4.717 1.63e-05 ***
## road         -2.4097     0.4365  -5.520 9.04e-07 ***
## weather       4.2456     0.4473   9.492 2.92e-13 ***
## fuel         -3.6145     2.2759  -1.588    0.118
## wind         -0.1358     0.1764  -0.769    0.445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.913 on 56 degrees of freedom
## Multiple R-squared:  0.7405, Adjusted R-squared:  0.7174
## F-statistic: 31.96 on 5 and 56 DF,  p-value: 3.039e-15
```

- ***wind*** has the highest P-value so remove it first

```r
M2 = update(M1, . ~ . - wind)
summary(M2)
```

```
##
## Call:
## lm(formula = spi ~ transport + road + weather + fuel, data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9347  -4.2440   0.0528   5.0544  21.4515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  61.1610     7.0669   8.655 5.69e-12 ***
## transport    -2.1257     0.4550  -4.672 1.86e-05 ***
## road         -2.4372     0.4335  -5.622 5.92e-07 ***
## weather       4.2565     0.4454   9.555 1.94e-13 ***
## fuel         -3.6853     2.2659  -1.626    0.109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.877 on 57 degrees of freedom
## Multiple R-squared:  0.7378, Adjusted R-squared:  0.7194
## F-statistic: 40.09 on 4 and 57 DF,  p-value: 5.959e-16
```

- ***fuel*** still has large P-value so remove it

```r
M3 = update(M2, . ~ . - fuel)
summary(M3)
```

```
##
## Call:
## lm(formula = spi ~ transport + road + weather, data = traffic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -21.672  -5.643   1.067   4.656  23.164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.7370     4.1027  12.611  < 2e-16 ***
## transport    -2.3216     0.4449  -5.218 2.54e-06 ***
## road         -2.4563     0.4394  -5.590 6.40e-07 ***
## weather       4.1450     0.4463   9.286 4.48e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.02 on 58 degrees of freedom
## Multiple R-squared:  0.7256, Adjusted R-squared:  0.7114
## F-statistic: 51.12 on 3 and 58 DF,  p-value: 2.724e-16
```

- In model M3 although all predictors are significant, the value of $R^2$ adjust decrease, which indicates that the recently removed predictor ***fuel*** has contribution to the model. Therefore, we will stop at the final model M2.

$$\hat{Y} = 61.16096 - 2.12565X_1 - 2.43721X_2 + 4.25645X_3 - 3.68533X_4$$

#### g) Explain R2 and R2 adjust:

- The R2 goodness of fit metric always decreases/increases when a predictor is removed/added from/into the model. The adjusted R2 has a penalty for the number of predictors in the model. So it will sometimes increase when a predictor is removed. In this case, from the full to final model, the R2 decreases from 74.1% to 73.8% but the adjusted R2 increases from 71.7% to 71.9%.

### Study on Cake dataset:

#### a) Balanced study checking and explain:

- A study is balanced if there are equal number of replicates across all the levels factors in the study.
- We can check replicates by:

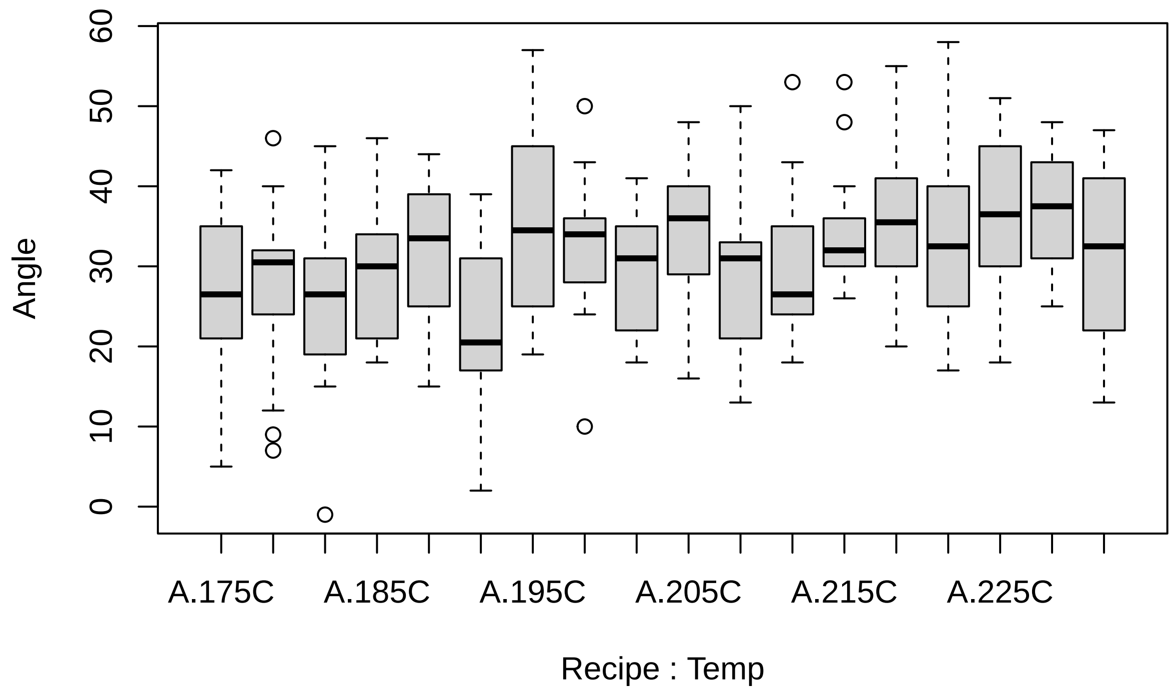
```r
cake  = read.csv('data/cake.csv', header = TRUE, stringsAsFactors = TRUE)
table(cake[, c("Recipe", "Temp")])
```

```
##       Temp
## Recipe 175C 185C 195C 205C 215C 225C
##      A   14   14   14   14   14   14
##      B   14   14   14   14   14   14
##      C   14   14   14   14   14   14
```

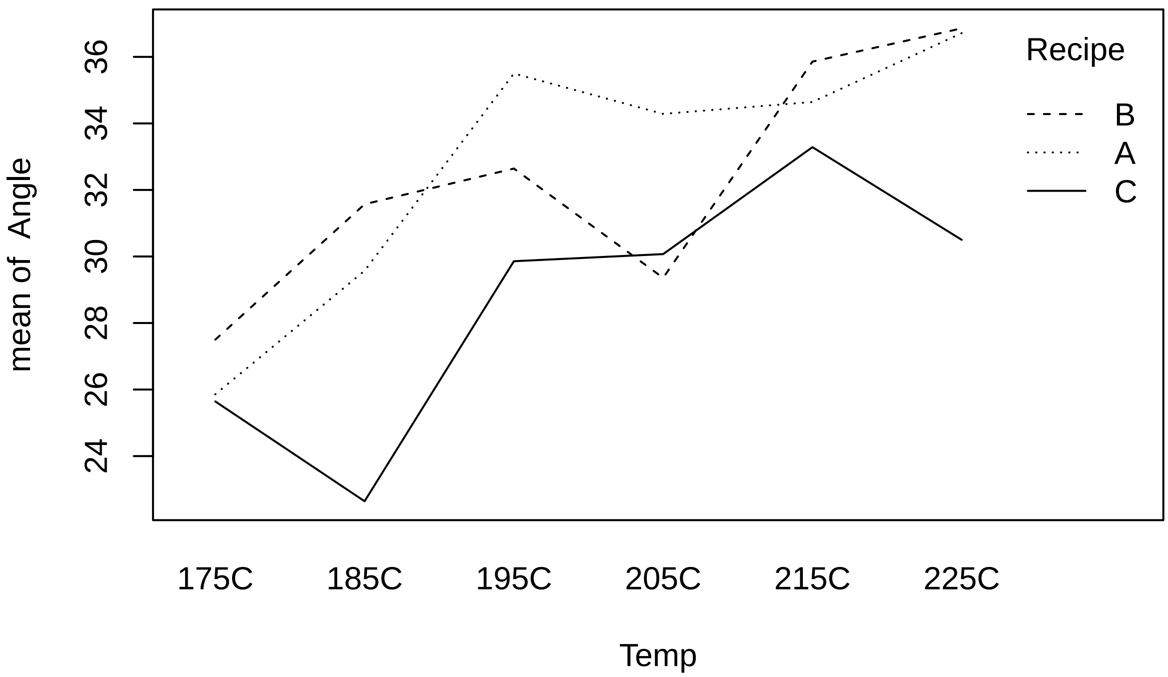
- From the above we can see that the design is balanced with an equal number of replicates for each combination of levels of the two factors.

#### b) Preliminary graphs that investigate different features of the data:

```r
boxplot(Angle ~ Recipe + Temp, data = cake)
```




```r
with(cake, interaction.plot(Temp, Recipe, Angle))
```




- From the boxplot, we can see that the assumption of equal variance among levels seems approximately valid due to the similar box sizes.
- From the interaction plot we can see there is non-parallel lines for the means of each group at different levels of the independent variables, this indicates a significant interaction effect between the two independent variables.

**c) Interaction model:**

- The Two-Way ANOVA model with interaction is:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

- The parameters are:
  - $Y_{ijk}$: the reponse breaking angle of the cake
  - $\alpha_i$: the ***Recipe*** effect, there are 3 levels: A, B, C
  - $\beta_j$: the ***Temp*** effect, there are 6 levels: 175C, 185C, 195C, 205C, 215C, 225C
  - $\gamma_{ij}$: the interaction effect between ***Recipe*** and ***Temp***
  - $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$: the unexplained variation

**d) Study the effect of *Recipe* and *Temp* on *Angle*:**

- We will conduct a hypotheses test:

$$H_0: \gamma_{ij} = 0 \text{ for all } i, j H_1: \text{at least one } \gamma_{ij} \neq 0$$

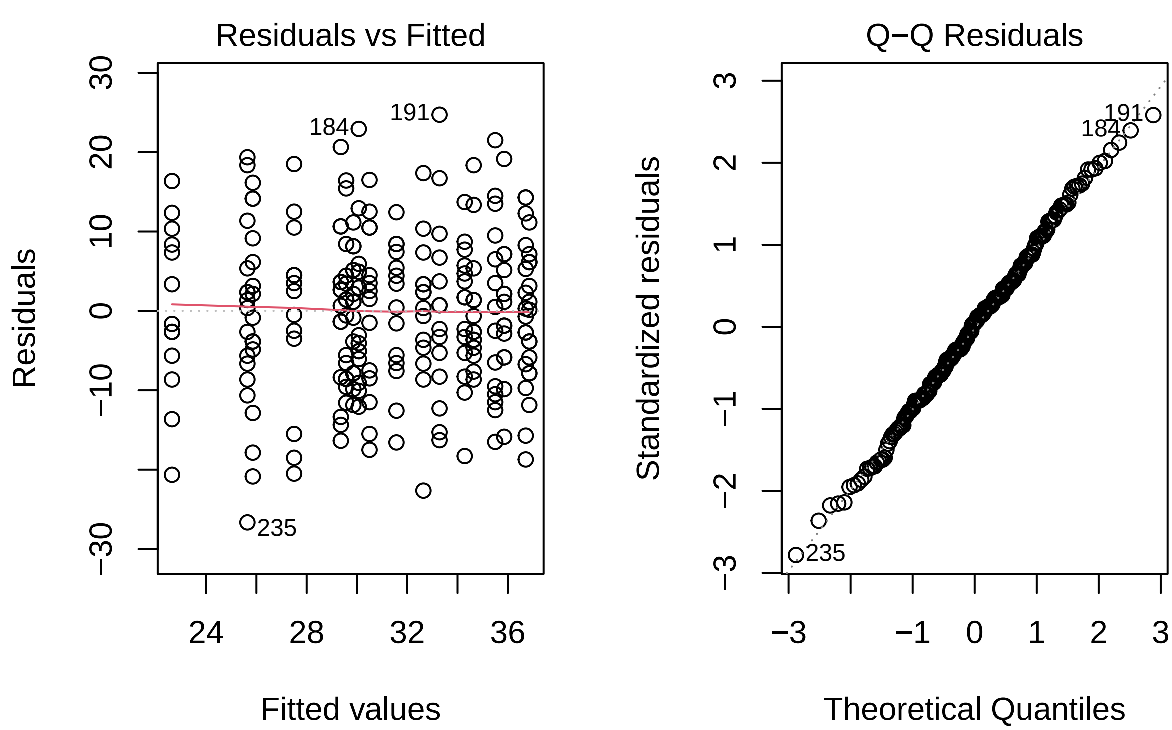
- Now we fit the interaction model:

```
cake.int = lm(Angle ~ Recipe * Temp, data = cake)
anova(cake.int)
```

```
## Analysis of Variance Table
##
## Response: Angle
##               Df  Sum Sq Mean Sq F value   Pr(>F)
## Recipe         2   844.8  422.38  4.2762 0.014998 *
## Temp           5  2530.1  506.01  5.1228 0.000177 ***
## Recipe:Temp   10   635.6   63.56  0.6435 0.775632
## Residuals    234 23113.8   98.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Since the P-value is $0.775632 > 0.05$. we do not have enough evidence to reject $H_0$. We can see that the interaction term is insignificant. Therefore, this is not the final model yet.
- We should validate the interaction model with diagnostic plots:

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



- The residuals are close to linear in the QQ-plot, and so the normal assumption should be valid. The residual plot seems to show equal spread around the fitted values and so the constant variance assumption is also appropriate.

**e) Repeat test analysis for the main effects:**

- We will conduct hypotheses tests for main effects:

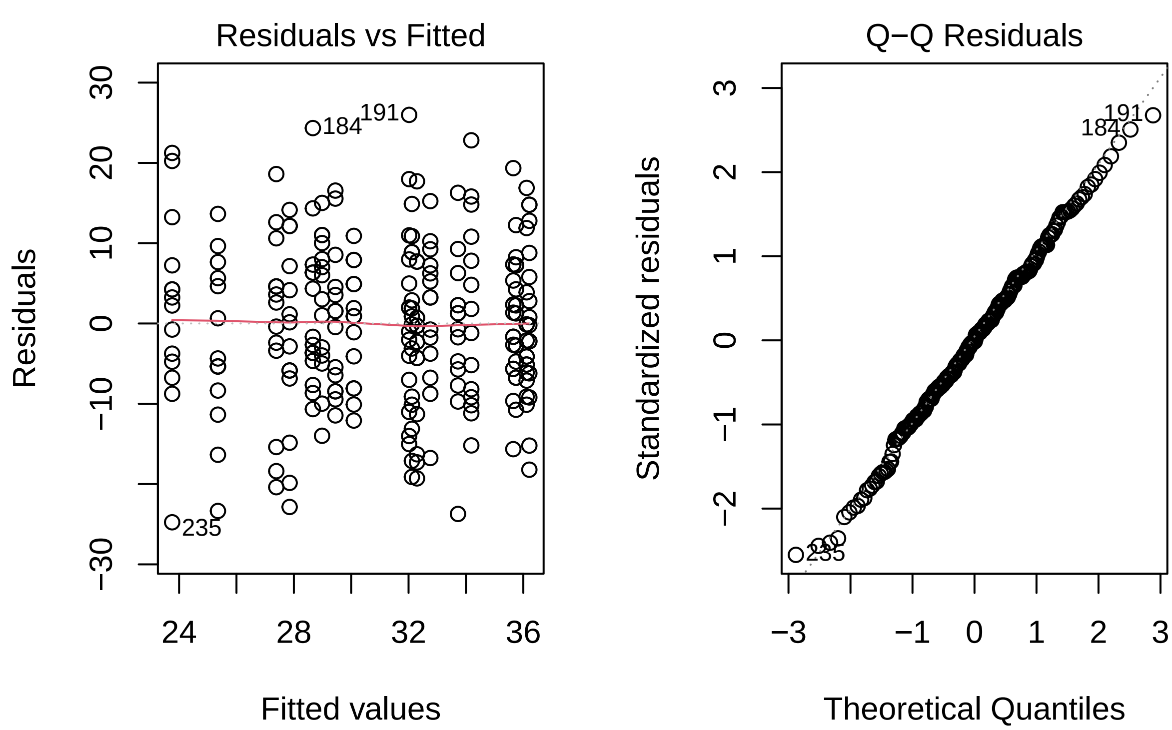
$$H_0 : \alpha_i = 0 \text{ for all } i \qquad \text{against} \qquad H_1 : \alpha_i \neq 0 H_0 : \beta_j = 0 \text{ for all } j \qquad \text{against} \qquad H_1 : \beta_j \neq 0$$

- Now we refit the model without interaction term:

```
cake.main = lm(Angle ~ Recipe + Temp, data = cake)
anova(cake.main)
```

```
## Analysis of Variance Table
##
## Response: Angle
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Recipe      2   844.8  422.38  4.3396 0.0140636 *
## Temp        5  2530.1  506.01  5.1988 0.0001489 ***
## Residuals 244 23749.4   97.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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



- The result of ANOVA table shows that the P-values for both ***Recipe*** and ***Temp*** are smaller than 0.05, which help reject all $H_0$ above and indicates that the main effects are significant.
- Again, the residuals are close to linear in the QQ-plot, and so the normal assumption should be valid, and the residual plot seems to show equal spread around the fitted values and so the constant variance assumption is also appropriate.

**f) Conclusion on effect:**

- Overall, the effect of ***Recipe*** and ***Temp*** on the quality of cakes ***Angle*** are not depend on each other since the interaction term is insignificant. However, these effects separately have significant impact on the response ***Angle***, that is, ***Angle*** seems to be higher with ***Recipe*** A and B, within the ***Temp*** ranges of 195C, 215C, 225C.
- We can interpret the main effects separately since there is no significant interaction effect them.