STAT2170 Assignment

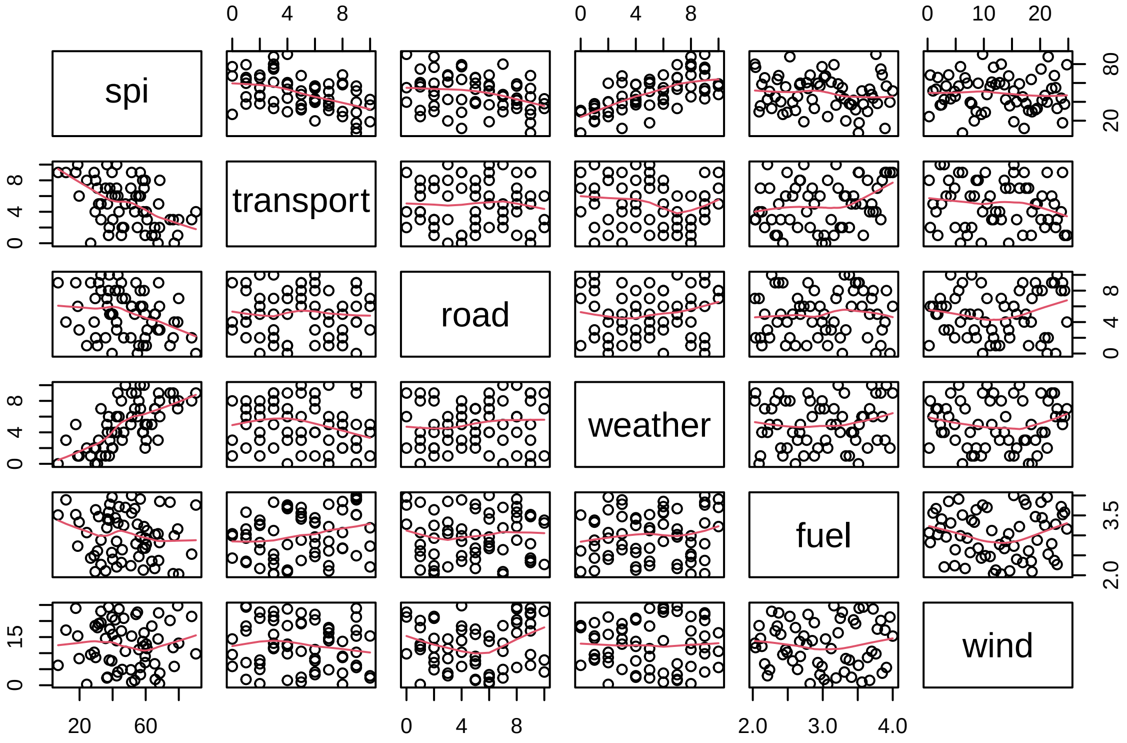
Benyamin Mahmoudkhalesi 46784365

2023-10-18

Question 1

1. Imporrting, Ploting and Correlation Matrix of Data

```r
traffic_data <- read.csv("data/traffic.csv", header = TRUE)
pairs(traffic_data, panel = panel.smooth)
```



## 1.1 Relationships between the response (`spi`) and predictors: - **spi vs. transport**: There appears to be a negative linear relationship between `spi` and `transport`. The correlation coefficient is approximately -0.473, suggesting a moderate negative correlation. - **spi vs. road**: There's a weak negative relationship with a correlation coefficient of about -0.304. - **spi vs. weather**: This pair shows a strong positive linear relationship with a correlation coefficient of approximately 0.667. - **spi vs. fuel**: A very weak negative

relationship is observed (correlation approximately −0.138). - **spi vs. wind**: The relationship is almost non-existent with a correlation close to zero.

```r
cor(traffic_data)
```

```
##                   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
```

1.2 Relationships between the predictors themselves:

Transport

- **vs. Road**: Very weak correlation, almost non-existent.
- **vs. Weather**: A weak negative relationship with a correlation of about −0.170.
- **vs. Fuel**: A weak positive relationship (correlation approximately 0.241).
- **vs. Wind**: Weak negative relationship with a correlation of −0.131.

Road

- **vs. Weather**: Weak positive relationship.
- **vs. Fuel** and **vs. Wind**: Both have very weak relationships.

Weather

- **vs. Fuel**: Weak positive correlation.
- **vs. Wind**: Almost no correlation.

Fuel vs. Wind

- Almost no correlation.

1.3 Conclusion:

- `spi` seems to be most strongly influenced by `weather`, followed by `transport`.
- Among predictors, there aren't any pairs with strong correlations, which is good from a multicollinearity perspective in multiple regression.

2. Fiting our model

```
traffic.lm <- lm(spi ~ ., data = traffic_data)
summary(traffic.lm)
```

```
##
## Call:
## lm(formula = spi ~ ., data = traffic_data)
##
## 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
```

```
coefficients(traffic.lm)
```

```
## (Intercept)   transport        road     weather        fuel        wind
##  62.8071258  -2.1750456  -2.4096862   4.2456436  -3.6145270  -0.1357686
```

Estimating the impact of weather on spi using a 95% confidence interval follows this formula:

$$\beta_{\text{weather}} \pm t_{0.975,56} \times \text{s.e.}(\beta_{\text{weather}})$$

$$\beta_{\text{weather}} = 4.2456$$
$$\text{s.e.}(\beta_{\text{weather}}) = 0.4473$$

```
t <- qt(0.975, 56)
Upperbound <- 4.2456 + t * 0.4473
Lowerbound <- 4.2456 - t * 0.4473
Upperbound
```

```
## [1] 5.14165
```

```
Lowerbound
```

```
## [1] 3.34955
```

The confidence interval does not contain zero, meaning that the effect of weather on spi is statistically significant at the 95% confidence level. For every unit increase in the weather index, the spi will increase by an amount between 3.34955 and 5.14165 units, holding all other predictors constant.

3. Mathematical Multiple Regression Model:

The mathematical model for multiple regression is:

$$\text{spi} = \beta_0 + \beta_1 \times \text{transport} + \beta_2 \times \text{road} + \beta_3 \times \text{weather} + \beta_4 \times \text{fuel} + \beta_5 \times \text{wind} + \epsilon$$

Where: β_0 is the intercept. $\beta_1, \beta_2, \ldots, \beta_5$ are the coefficients for the predictors. ϵ is the error term.

3.1 Hypotheses for the Overall ANOVA test of Multiple Regression:

The null and alternative hypotheses are: $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (No relationship between predictors and response) H_a: At least one $\beta_i \neq 0$ (There is a relationship between at least one predictor and the response)

3.2 ANOVA table for the overall multiple regression model

```
anova_table <- anova(traffic.lm)
print(anova_table)
```

```
## 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
```

$$F = \frac{\left(\text{SS}_{\text{transport}} + \text{SS}_{\text{road}} + \text{SS}_{\text{weather}} + \text{SS}_{\text{fuel}} + \text{SS}_{\text{wind}}\right)/5}{\text{MS}_{\text{error}}}$$

3.3 Null distribution for the test statistic:

The null distribution for the F-test in multiple regression is the F-distribution with p and n-p-1 degrees of freedom where p is the number of predictors and n is the number of observations. p = 5 **(because there are 5 predictors: transport, road, weather, fuel, and wind)** and n = 62 the null distribution corresponds to $F_{5,56}$

3.4 Computing P-value and F statistic:

The p-value represents the probability of observing a test statistic as extreme as, or more extreme than, the statistic calculated from the sample data, assuming the null hypothesis is true. The P value is $1 - F_{\text{CDF}}(F, df_1, df_2)$

```
# Extracting sum of squares from the ANOVA table
SS_transport <- 4742.6
SS_road <- 1992.7
SS_weather <- 8651.9
SS_fuel <- 258.1
SS_wind <- 58.2

# Mean squared error
MS_error <- 98.3

# Compute the F statistic
F = (SS_transport + SS_road + SS_weather + SS_fuel + SS_wind) / 5 / MS_error

# Compute the p-value for the F statistic
p_value = 1 - pf(F, 5, 56)

F
```

```
## [1] 31.95015
```

```
p_value
```

```
## [1] 3.108624e-15
```

The F-statistic is 31.95015 and P-value is 3.108624e-15

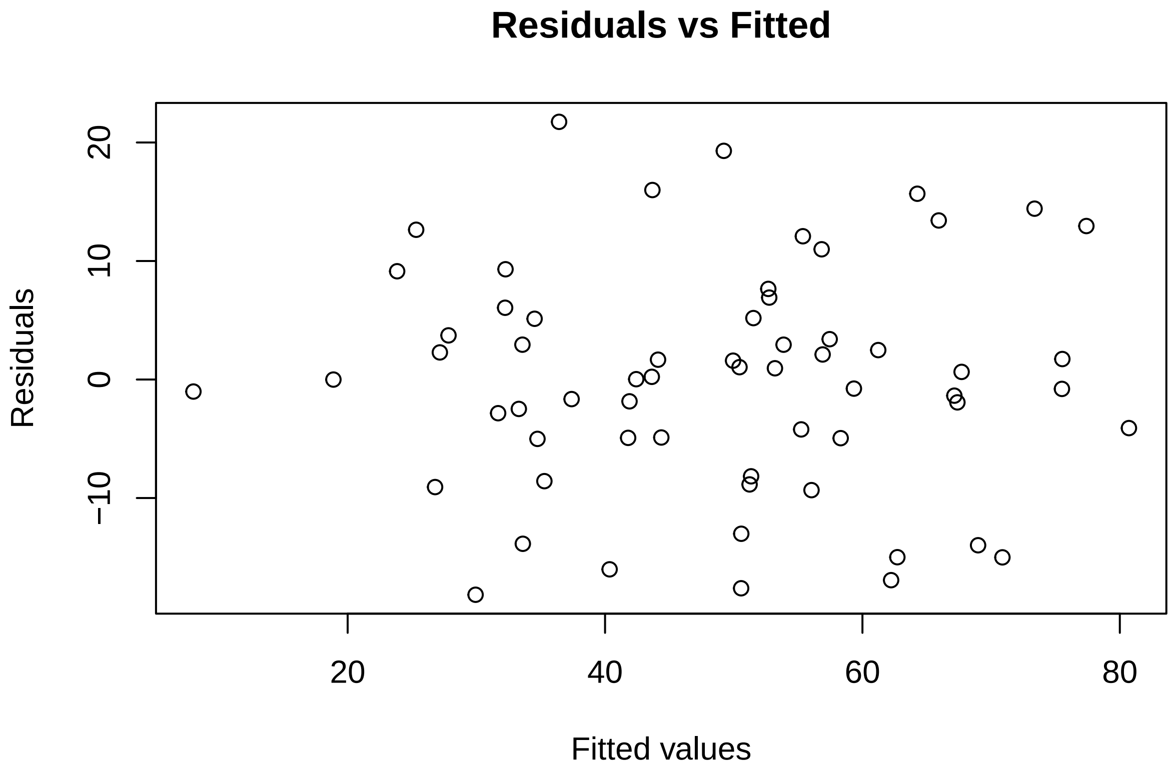
3.5 Conclusion:

The extremely small p-value (much less than 0.05) provides strong evidence against the null hypothesis. Therefore, we can reject the null hypothesis, In other words, there’s a significant relationship between the response variable and at least one of the predictor variables in the model.

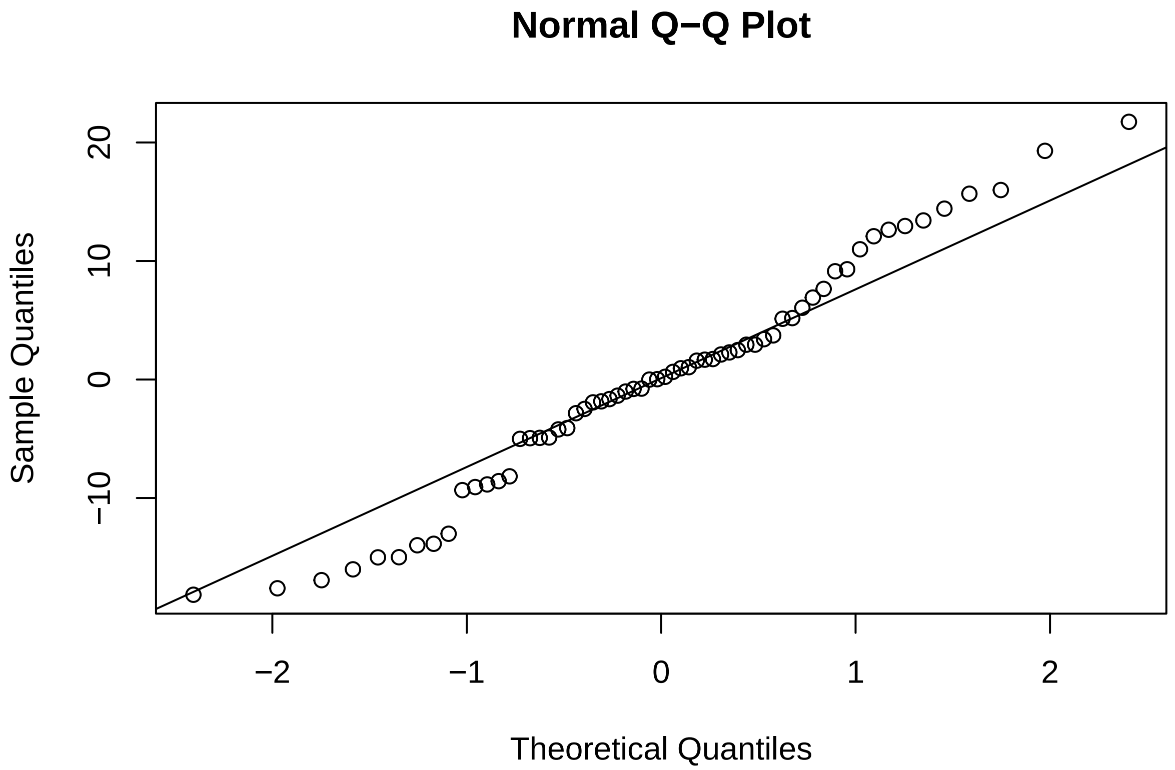
4. Validation & Diagnostics

To validate the full regression model and determine whether it is appropriate to explain spi we need to run our diagnostic measures:

```
plot(traffic.lm$fitted.values, traffic.lm$residuals, main="Residuals vs Fitted", xlab="Fitted values", 
```



```
qqnorm(traffic.lm$residuals, main="Normal Q-Q Plot")
qqline(traffic.lm$residuals)
```



Residuals vs. fitted plot shows a random scatter without any pattern thus he assumptions of linearity and equal variance are met. The Q-Q plot also shows that the points lie roughly along the reference line, the assumption of normality is met. **the model is appropriate.**

4.1 R^2

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

R^2represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides a measure of how well the observed outcomes are replicated by the model. Based on the model summary 74.05% of the variability in the response spi is explained by the predictors in the model. This is a relatively high R^2 value, suggesting that our **model fits the data quite well.**

5. Finding the best model

Based on the previous fitting The predictor wind has the largest p-value of approximately 0.445 which is greater than the threshold of 0.05, so I will remove the wind predictor and refit the model.

```
traffic.lm2 <- lm(spi ~ transport + road + weather + fuel, data = traffic_data)
summary(traffic.lm2)
```

```
##
## Call:
```

```
## lm(formula = spi ~ transport + road + weather + fuel, data = traffic_data)
##
## 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
```

After removing the wind predictor and refitting the model: - The predictor fuel has the largest p-value of approximately 0.109 which is greater than the significance threshold of 0.05 so I will again remove the fuel predictor and refit the model.

```
traffic.lm3 <- lm(spi ~ transport + road + weather, data = traffic_data)
summary(traffic.lm3)
```

```
##
## Call:
## lm(formula = spi ~ transport + road + weather, data = traffic_data)
##
## 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
```

After removing the fuel predictor and refitting the model: - All predictors in the model (transport, road, and weather) have p-values less than the significance threshold of 0.05 and will be utilized for our final best model.

```
coefficients(traffic.lm3)
```

```
## (Intercept)   transport        road     weather 
##   51.737015   -2.321620   -2.456284    4.144978
```

$$\text{spi} = 51.7370 - 2.3216 \times \text{transport} - 2.4563 \times \text{road} + 4.1450 \times \text{weather}$$

6. R^2 and Adjusted R^2 explanation

- **Original Full Model: Multiple R-squared:** 0.7405, **Adjusted R-squared:** 0.7174
- **Final Best Model: Multiple R-squared:** 0.7256, **Adjusted R-squared:** 0.7114

The R2 for the original full model is slightly higher than that of the final best model. This is expected because R2 will generally not decrease (and often increases) when new predictors are added, even if they don't have a significant relationship with the dependent variable.

6.2 Introducing Adjusted R2

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1}\right)$$

- The adjusted R^2 takes into account the number of predictors in the model. It increases only if the new predictor enhances the model more than would be expected by chance. It can decrease if the predictor doesn't improve the model significantly.
- The adjusted R^2 for the full model is also slightly higher than the final model. However, the difference between them is smaller than the difference in R^2 values. This shows that some predictors in the full model were not adding substantial value, as the penalty for including them brings the adjusted R^2 closer to the final model's value.

Question 2

1. Balanced or Unbalanced?

To determine if the design is balanced, we need to check the number of observations for each combination of Recipe and Temp

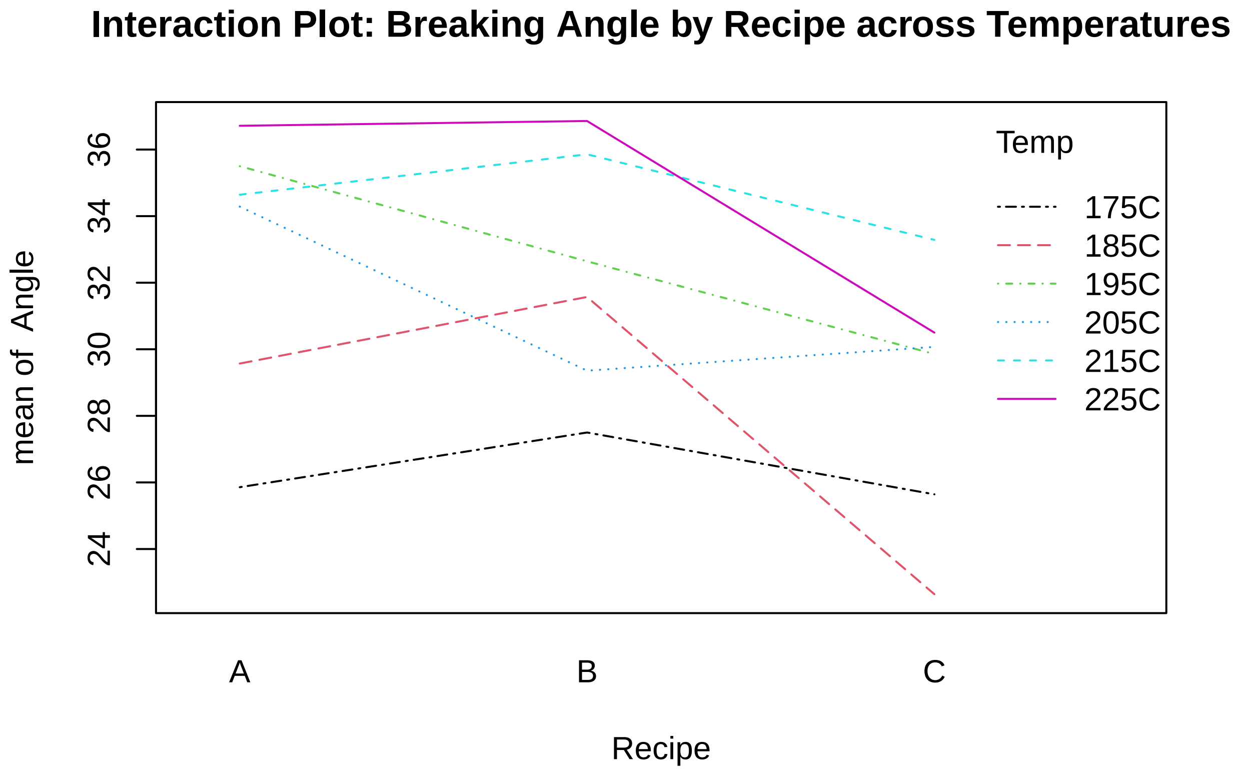
```
cake_data <- read.csv("data/cake.csv", header = TRUE)
table(cake_data[, c("Temp", "Recipe")])
```

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

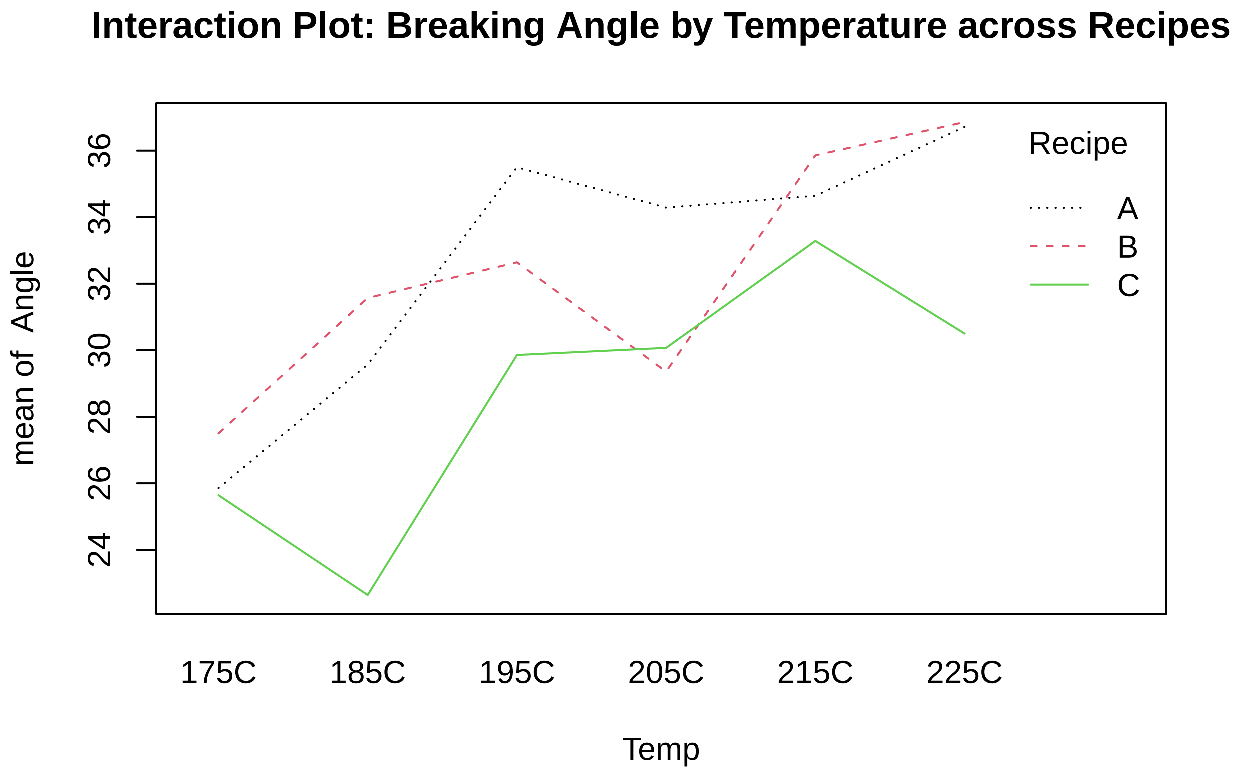
The data shows that each combination of Recipe and Temp has exactly 14 observations. Given this, the design of the study is balanced because each combination of factors (recipe and temperature) has an equal number of observations.

1.1 Preliminary graphs

```
with(cake_data, interaction.plot(Recipe, Temp, Angle, col = 1:6, fixed = TRUE, main = "Interaction Plot
```



```
with(cake_data, interaction.plot(Temp, Recipe, Angle, col = 1:6, fixed = TRUE, main = "Interaction Plot
```



## 1.2 Interaction Plot: Breaking Angle by Recipe across Temperatures

Observations: - The variability in breaking angles across temperatures is pronounced for some recipes (like Recipe B) and relatively consistent for others (like Recipe C). - Recipe B at 175°C and 185°C has notably higher breaking angles compared to other recipes at the same temperatures. **- The lines for different temperatures are not parallel, indicating interaction effects between recipes and temperatures. This means that the effect of temperature on the breaking angle is not consistent across recipes.**

- Recipe A: The breaking angles seem to be relatively consistent across temperatures, with a slight dip at 195°C. The range of breaking angles for this recipe is moderate, suggesting that Recipe A produces cakes with consistent quality across the given temperature range.
- Recipe B: This recipe displays the most variability across temperatures. At 175°C and 185°C, the breaking angles are significantly higher than those for other recipes at the same temperatures. However, as the temperature increases to 195°C and beyond, the breaking angles decrease noticeably. This suggests that Recipe B might be more sensitive to temperature changes.
- Recipe C: The breaking angles for this recipe are relatively consistent across temperatures, with only minor fluctuations. This indicates that Recipe C might be more robust or less sensitive to temperature variations compared to other recipes.

2. Interaction Model:

$$\text{Angle} = \mu + \text{Temp}_i + \text{Recipe}_j + (\text{Temp} \times \text{Recipe})_{ij} + \epsilon$$

2.1 Hypotheses:

Interaction Effect (H_{01}): H_0: There's no interaction between temperature and recipe regarding the breaking angle. $H_0 : \gamma_{ij} = 0$ for all i, j; H_a: There's at least one significant interaction between temperature and recipe. H_a: not all $\gamma_{ij} = 0$.

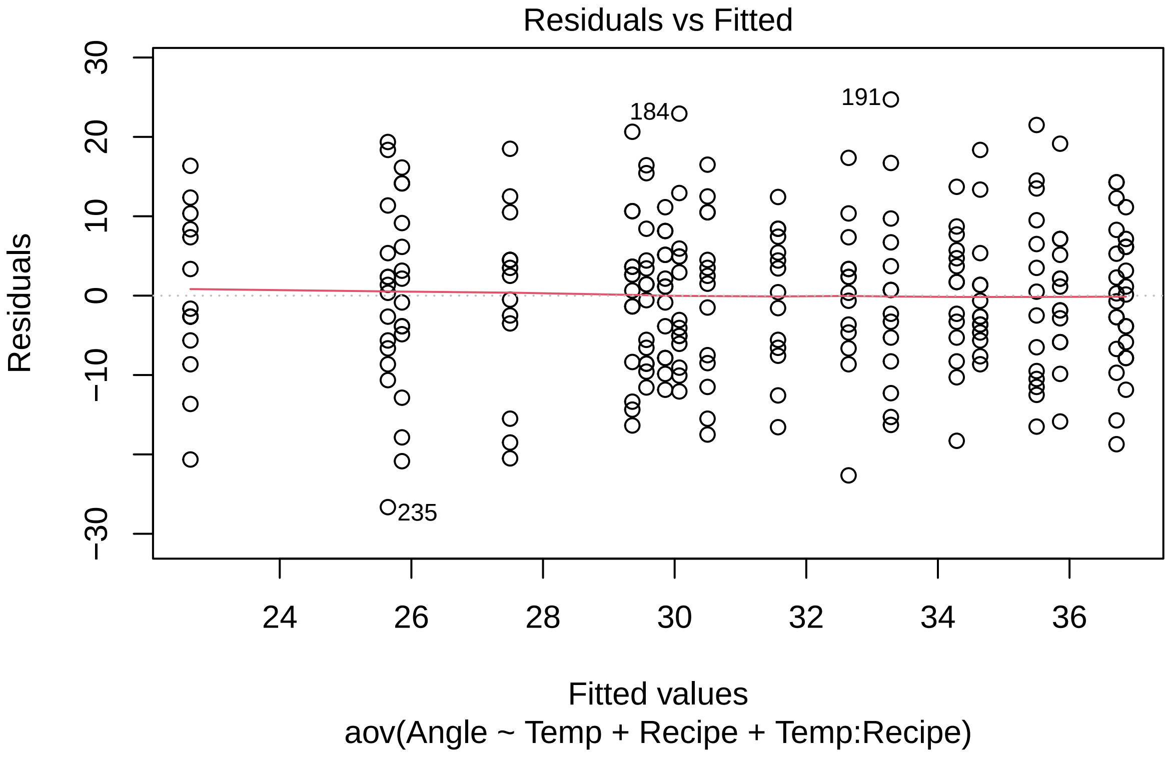
```
Anova_model <- aov(Angle ~ Temp + Recipe + Temp:Recipe, data=cake_data)
summary(Anova_model)
```

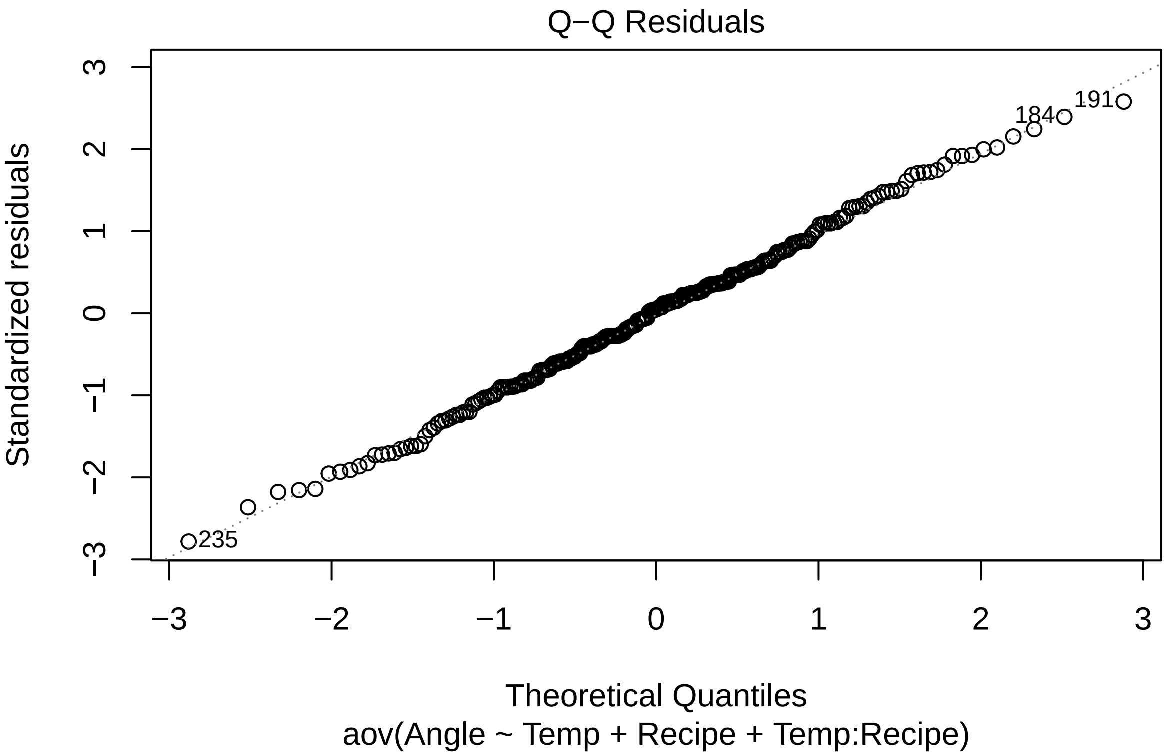
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Temp          5   2530   506.0   5.123 0.000177 ***
## Recipe        2    845   422.4   4.276 0.014998 *
## Temp:Recipe  10    636    63.6   0.643 0.775632
## Residuals   234  23114    98.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction Effect: The p-value was 0.775632, which is greater than 0.05. Thus, we fail to reject the null hypothesis H_{01}, suggesting that there's no significant interaction effect between temperature and recipe on the breaking angle. we do the subsequent tests but before that we do a diagnostic check.

2.2 Model Diagnostics

```
plot(Anova_model, which = 1:2)
```





both plots suggest that the key assumptions for ANOVA (normality of residuals and equal variances) are reasonably satisfied. he residuals in our plot seem to be randomly scattered around the horizontal line, without any clear funnel shape or curve, which is a good sign and the assumptions of linearity and homoscedasticity equal variances are met. the QQ plot suggests that the residuals are roughly normally distributed as the majority of the data points lie close to the line.

Temperature Effect (H_{02}): H_0: The mean breaking angle is the same across all temperature levels. H_a: At least one temperature level has a different mean breaking angle.

Based on the Anova table P-value is 0.000177 which is less than 0.05. we can reject the H_0 and this suggests that the temperature at which the cake was baked has a significant effect on the breaking angle.

Recipe Effect (H_{03}): H_0: The mean breaking angle is the same across all recipe types. H_a: At least one recipe type has a different mean breaking angle.

Based on the Anova table P-value is 0.014998 which is less than 0.05. we can reject the H_0 and this suggests that the recipe used also has a significant effect on the breaking angle.

3. Main Effects Model:

$$\text{Angle} = \mu + \text{Temp}_i + \text{Recipe}_j + \epsilon$$

```
model_main_effects <- aov(Angle ~ Temp + Recipe, data = cake_data)
summary(model_main_effects)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## Temp            5   2530   506.0   5.199 0.000149 ***
## Recipe        2    845   422.4   4.340 0.014064 *
## Residuals   244  23749    97.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Temp (Temperature):

- **F value:** 5.199
- **p value:** 0.000149

The p-value is less than the typical significance level of 0.05. This indicates that the temperature at which the cake was baked has a statistically significant effect on the breaking angle.

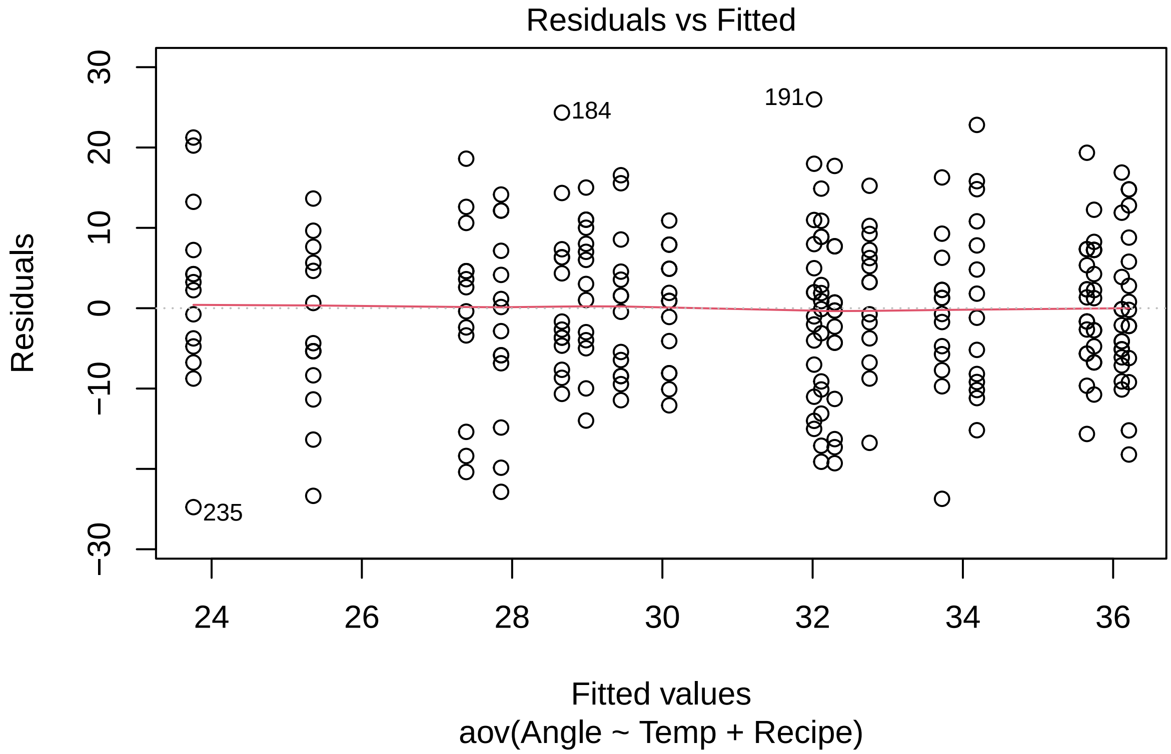
Recipe:

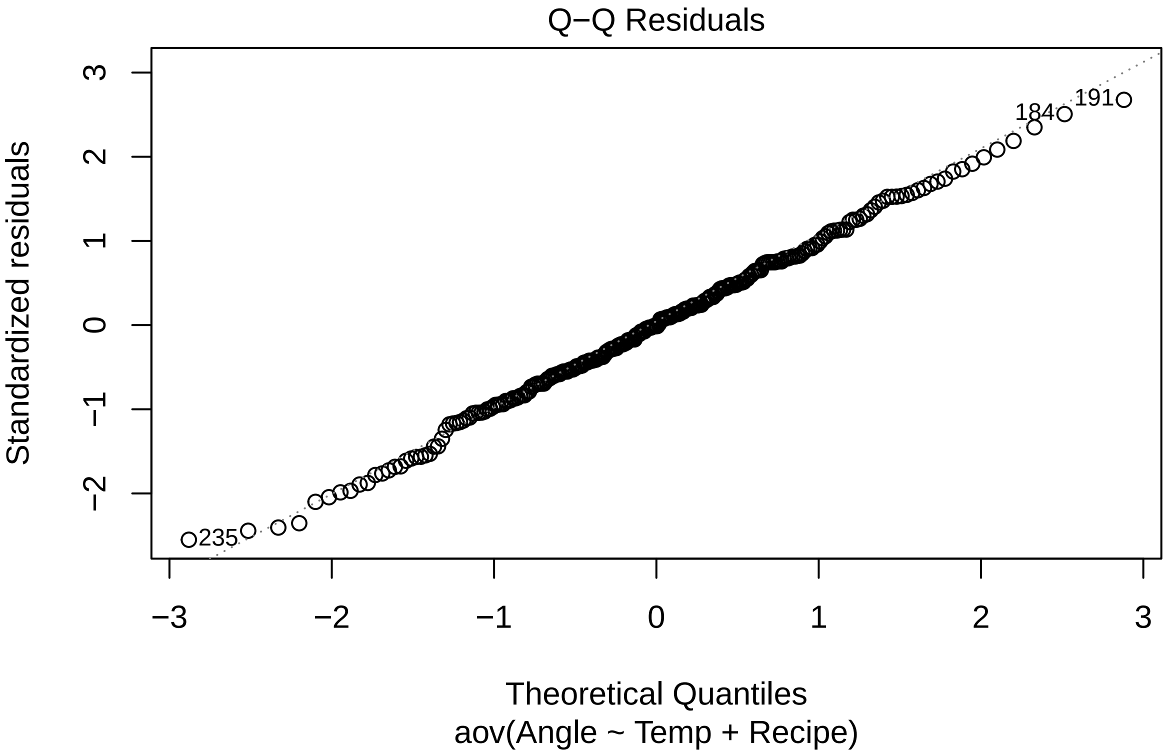
- **F value:** 4.340
- **p value:** 0.014064

The p-value is also less than 0.05, suggesting that the recipe used for the cake has a statistically significant effect on the breaking angle.

3.1 Diagnostics:

```
plot(model_main_effects, which = 1:2)
```





# 4.Conclusions: Temperature: There’s a statistically significant difference in the breaking angle based on the temperature at which the cake was baked. This implies that the baking temperature plays a crucial role in determining the quality of the cake, as measured by the breaking angle.

Recipe: Different recipes also lead to statistically significant differences in the breaking angle of the cake. This indicates that the ingredients and the method used in the recipe can influence the cake’s quality.