STAT2170 Assessment

2023-10-15

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

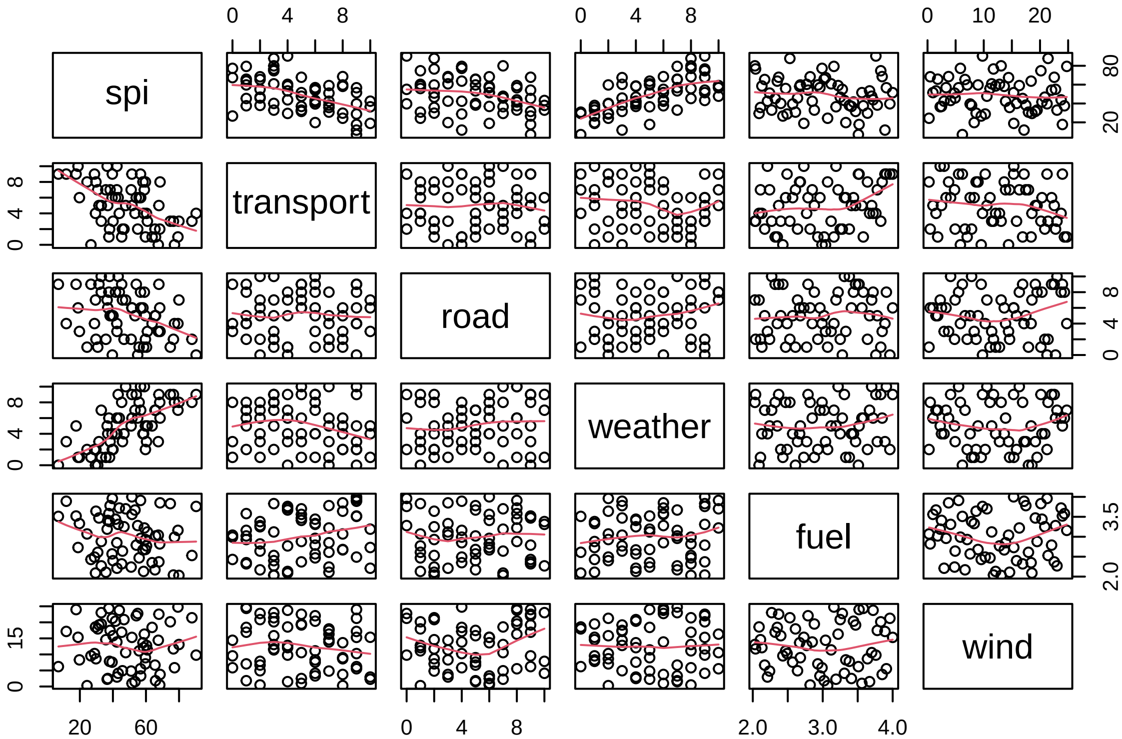
```r
#Question 1a

traffic = read.csv("traffic.csv", header = TRUE)

str (traffic)
```

```
## 'data.frame':    62 obs. of  6 variables:
##  $ spi      : num  39.5 36.9 47.7 58.2 60.3 ...
##  $ transport: int  6 5 5 8 4 3 4 4 9 8 ...
##  $ road     : int  5 6 7 6 1 2 5 1 8 3 ...
##  $ weather  : int  4 4 10 4 3 9 7 4 10 6 ...
##  $ fuel     : num  2.57 3.41 3.7 2.67 2.77 2.04 3.67 2.13 3.91 3.85 ...
##  $ wind     : num  7.58 2.49 10.56 13.64 12.8 ...
```

```r
head (traffic)
```

```
##     spi transport road weather fuel  wind
## 1 39.48         6    5       4 2.57  7.58
## 2 36.87         5    6       4 3.41  2.49
## 3 47.72         5    7      10 3.70 10.56
## 4 58.17         8    6       4 2.67 13.64
## 5 60.33         4    1       3 2.77 12.80
## 6 76.61         3    2       9 2.04 11.73
```

```r
pairs(traffic, panel = panel.smooth)
```




```r
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
```

```r
#The predictors "transport," "road," "fuel," and "wind" have a slight negative linear association with
#spi has a fairly positive linear connection with the "weather" predictor, which means that when "weath
#The correlations calculated in the matrix correspond to what you see in scatterplots, confirming your
#There are no strong or evident correlations between the predictor variables, suggesting that they are
```

```r
#Question1b

#fit a model

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)

se <- sqrt(diag(summary.M1$cov.unscaled*
 summary.M1$sigma^2))[3]


#The confidence interval (CI) for the influence of the "weather" predictor on "spi" is determined as $h
#Substituting the values yields the CI of $(3.34956864, 5.14163136)$.
#This indicates that, with 95% certainty, for each unit rise in "weather," the change in "spi" is proje
```

Question 2c

$Y_i = \beta_0 + \beta_1 X_1 + \epsilon + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$

Y is the response variable *spi*;

$X_1 = transport$ $X_2 = road$ $X_3 = weather$ $X_4 = fuel$ $X_5 = wind$

ϵ defines the random variation with constant variance

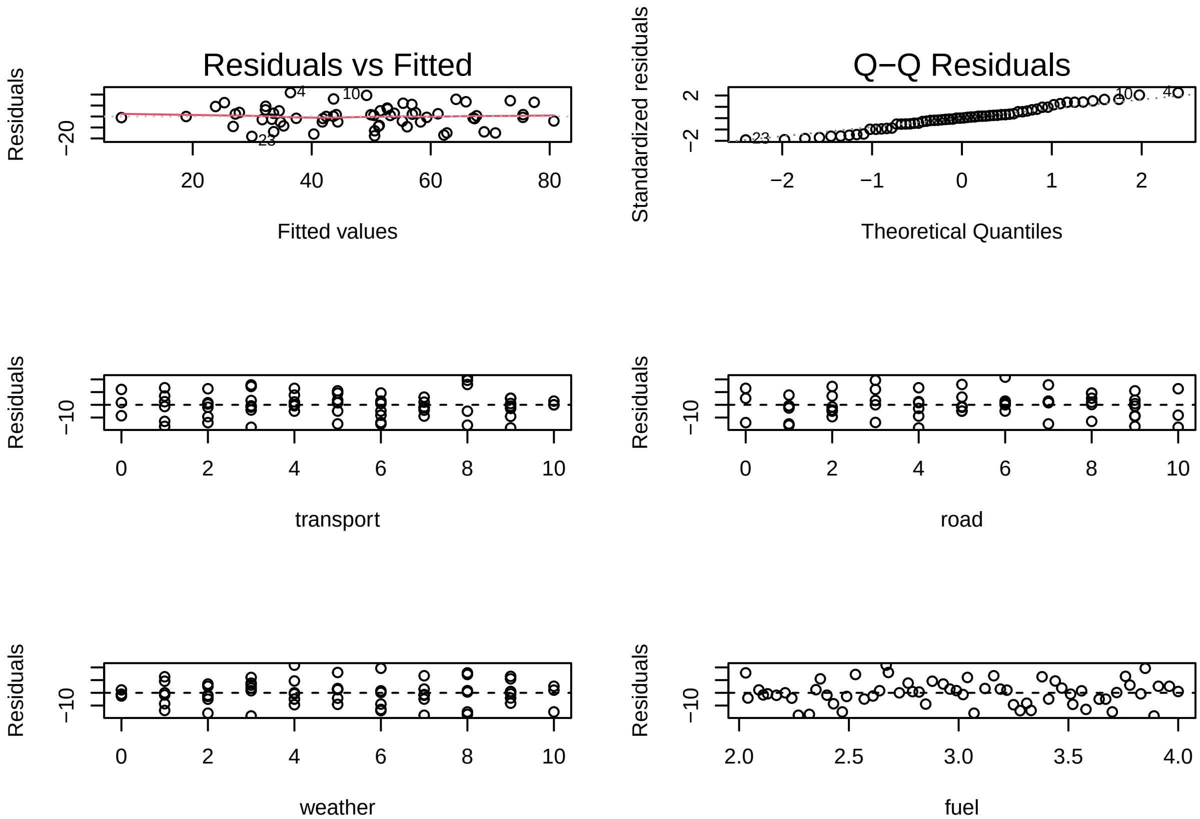
```r
anova(M1)
```

```
## 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
```

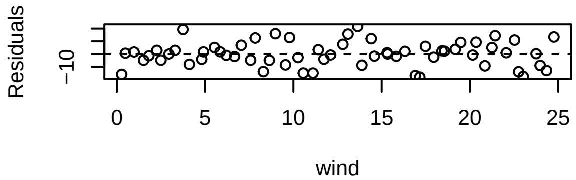
```r
#The total sum of squares (Regression SS) is determined by adding the sum of the individual components,
#Divide the Regression SS by the degrees of freedom (5) to get the mean square regression (Mean Square 
```

```
#The F-Statistic is determined as the mean square reg/mean square residual (MS_Res) ratio, which is rou
#The test statistic's null distribution is an F-distribution with degrees of freedom (5, 56).
#The p-value is calculated as P(F_5,56 31.9499), yielding an extremely tiny p-value of around $3.064 ti
#The null hypothesis ($H_0$) is rejected since the p-value is substantially lower than the significance
```

```
#Question 1d

par(mfrow = c(3, 2))
plot(M1, which = 1:2)
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)
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)

#There is no visible pattern in the residual plot, indicating that the residuals are randomly distribut
#Furthermore, the quantile plot of residuals reveals a linear trend. When the quantiles of the residual
#These findings show that the linear multiple regression model is acceptable for your data. The absence
```




```r
#Question 1e

summary(M1)$r.squared
```

```
## [1] 0.7405181
```

```r
#The coefficient of determination, given as $R2$, is determined to be roughly 0.7405181, or 74.05% when
#In a regression model, $R2$ is a measure of goodness of fit. It measures how effectively the linear re
#The $R2$ value of 0.7405181 (74.05%) suggests that the linear regression model accounts for 74.05% of
```

```r
# Question 1f
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
M2 <- update(M1, . ~ . - transport)

summary(M2)
```

```
##
## Call:
## lm(formula = spi ~ road + weather + fuel + wind, data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.0025  -7.8624  -0.5718   8.1327  23.6583
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 57.12442    8.56417   6.670 1.13e-08 ***
## road        -2.44965    0.51135  -4.791 1.23e-05 ***
## weather      4.67887    0.51289   9.123 9.74e-13 ***
## fuel        -6.48807    2.56933  -2.525   0.0144 *
## wind        -0.01989    0.20473  -0.097   0.9229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.61 on 57 degrees of freedom
## Multiple R-squared:  0.6374, Adjusted R-squared:  0.612
## F-statistic: 25.05 on 4 and 57 DF,  p-value: 5.317e-12
```

```
#All of the remaining predictor variables are statistically significant, indicating that they have a su
#$Y = 51.7370 - 2.3216X1 - 2.4563X2 + 4.1450X3' is the final model equation.
#This equation shows how the predictor variables "X1" (e.g., transport), "X2" (e.g., road), and "X3" (e
# The equation predicts "spi" as "spi = 51.7370 - 2.3216 * transport - 2.4563 * road + 4.1450 * weather
```

Question 1g

When a predictor is eliminated from a model, $R2$ often declines since it explains less variation. Adjusted $R2" takes into account the number of predictors; it may rise when a predictor is eliminated. The shift from the complete model to the final model results in minor declines in both $R2$ (from 74.05% to 72.56%) and adjusted $R2$ (from 71.7% to 71.1%). The most important finding, however, is that the p-value in the final model is significantly higher, indicating that it is a better, more parsimonious explanation for the data. The lower the p-value, the more relevant the predictors kept in the final model are in explaining the response variable.

```
## Question 2a

cake = read.csv("cake.csv", header = TRUE)
str(cake)
```

```
## 'data.frame':    252 obs. of  3 variables:
##  $ Temp  : chr  "185C" "175C" "195C" "205C" ...
##  $ Recipe: chr  "A" "A" "A" "A" ...
##  $ Angle : int  45 25 39 48 30 42 28 24 23 32 ...
```

```
head(cake)
```

```
##   Temp Recipe Angle
## 1 185C      A    45
## 2 175C      A    25
## 3 195C      A    39
## 4 205C      A    48
## 5 215C      A    30
## 6 225C      A    42
```

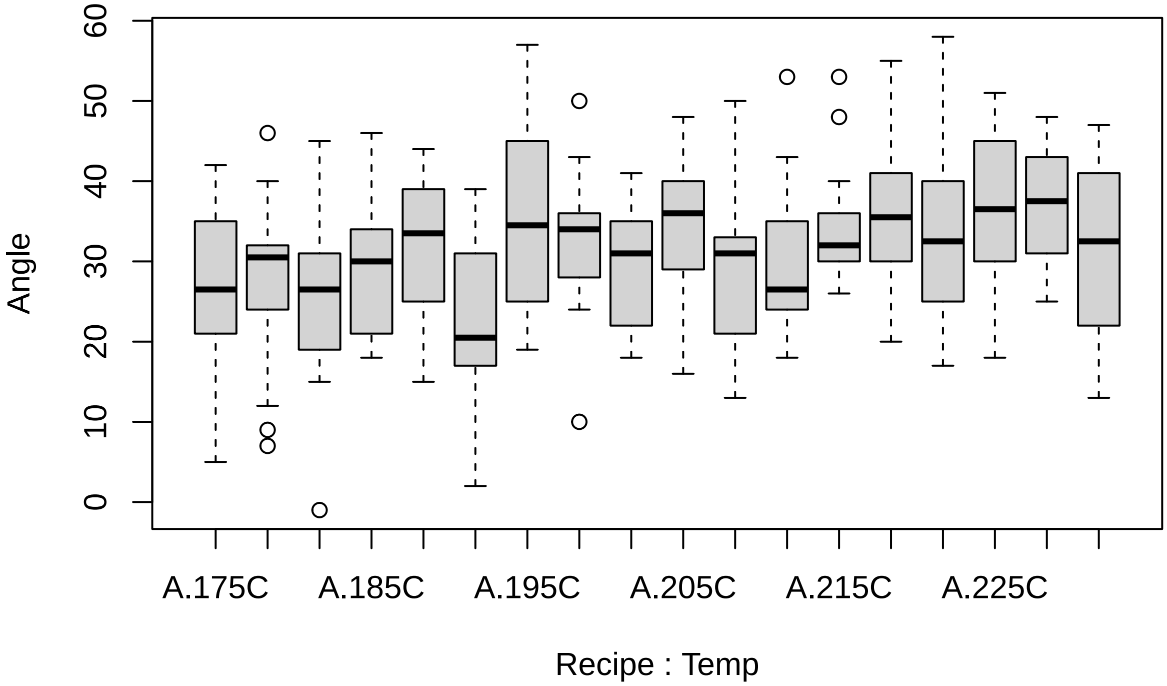
```
stringsAsFactors = TRUE

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
```

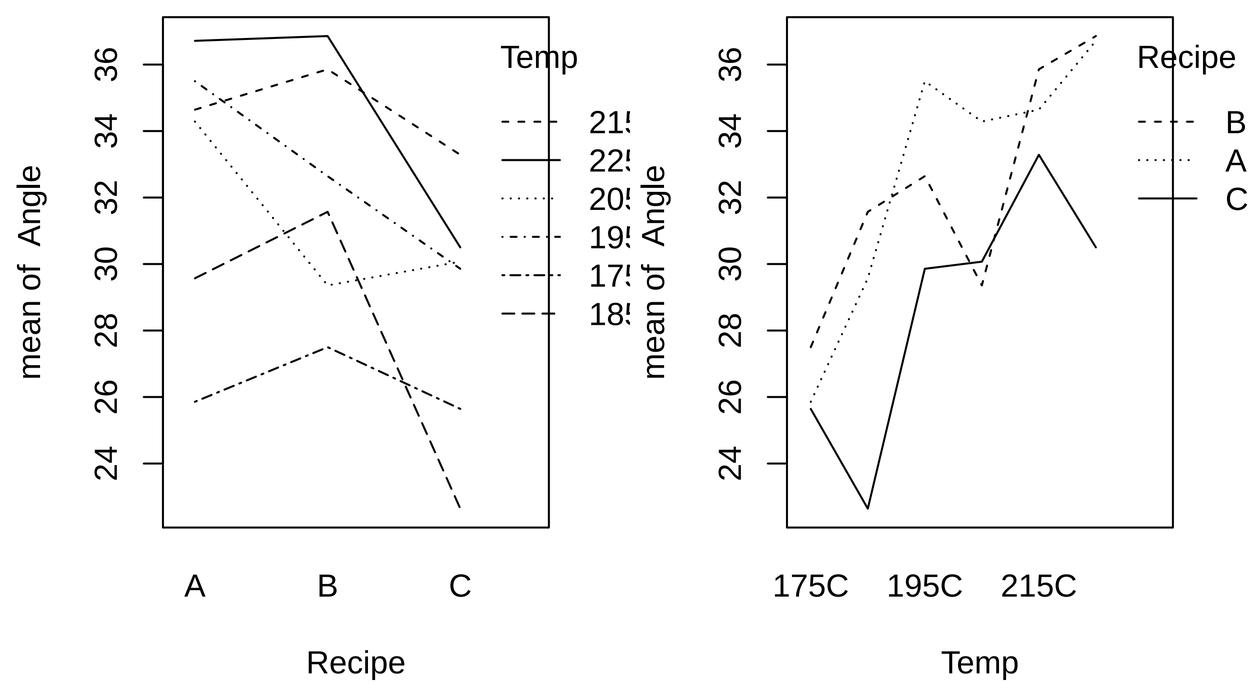
```
# The phrase "design" refers to the framework of an experiment or study, which comprises components, tr
# The study is defined as "balanced" because each condition or treatment contains an equal number of re
```

```r
# A balanced design is useful because it helps to guarantee that the results are not skewed by an unbal
```

```r
# Question 2b

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




```r
par(mfrow = c(1, 2))

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



```r
# The "interaction plot" shows how temperature changes affect different recipes differently and how the
# For example, increasing the temperature from 175°C to 185°C resulted in a comparable rise in the "ang
# These findings indicate that the effect of temperature on the angle is not constant across all recipe
```

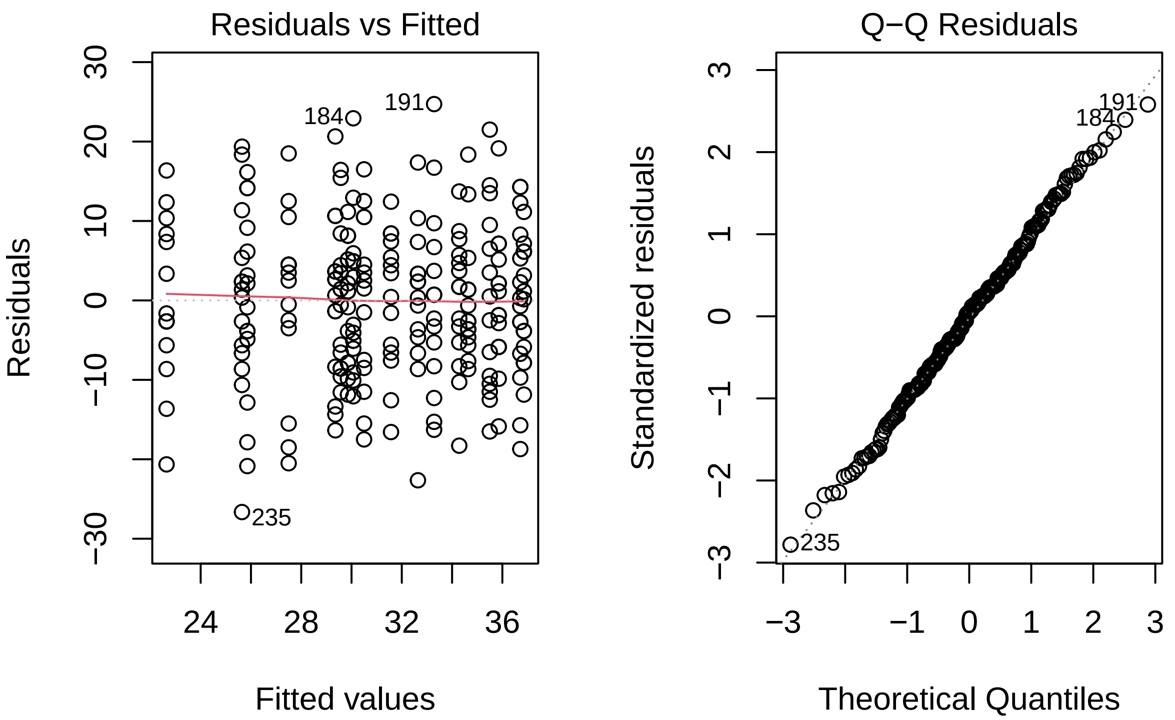
Question 2c

$Y = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon$ where Y is the breaking *angle* of cake' μ is the mean response α_i is the *temperature* main effect β_j is the *recipe* main effect γ_{ij} is the interaction effect between *temperature* and *recipe* and ϵ is the unexplained variation

```r
# Question 2d

cake.voc <- lm(Angle ~ Temp * Recipe, data = cake)
anova(cake.voc)
```

```
## Analysis of Variance Table
##
## Response: Angle
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## Temp          5  2530.1  506.01  5.1228 0.000177 ***
## Recipe        2   844.8  422.38  4.2762 0.014998 *
## Temp:Recipe  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
```

```r
# The interaction terms are found negligible by the F-test findings and can be eliminated. The diagnost

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




```r
#The residuals QQ-plot shows a linear trend with just slight asymmetry. This implies that the residuals
#The residual plot demonstrates a very uniform and consistent distribution of residuals around the expe
```

```r
#Question 2e

# $H_0 : \alpha_i = 0$ for all $i$ against $H_1$ : at least one $\alpha_i = 0$
# $H_0 : \beta_j = 0$ for all $j$ against $H_1$ : at least one $\beta_j = 0$

cake.test1 = lm(Angle ~ Recipe + Temp, data = cake)
summary(cake.test1)
```

```
##
## Call:
## lm(formula = Angle ~ Recipe + Temp, data = cake)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.7579  -6.2698  -0.1151   7.1706  25.9802
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.8532     1.7578  15.845  < 2e-16 ***
## RecipeB      -0.4643     1.5223  -0.305 0.760638
## RecipeC      -4.0952     1.5223  -2.690 0.007636 **
## Temp185C      1.5952     2.1529   0.741 0.459421
## Temp195C      6.3333     2.1529   2.942 0.003577 **
## Temp205C      4.9048     2.1529   2.278 0.023579 *
## Temp215C      8.2619     2.1529   3.838 0.000158 ***
## Temp225C      8.3571     2.1529   3.882 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.866 on 244 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.0993
## F-statistic: 4.953 on 7 and 244 DF,  p-value: 2.95e-05
```

```r
anova(cake.test1)
```

```
## 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
```

```r
cake.test2 = lm(Angle ~ Temp + Recipe, data = cake)
anova(cake.test2)
```

```
## Analysis of Variance Table
##
## Response: Angle
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Temp        5  2530.1  506.01  5.1988 0.0001489 ***
## Recipe      2   844.8  422.38  4.3396 0.0140636 *
## Residuals 244 23749.4   97.33
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#For main effects, hypothesis testing is performed. It determines whether "Recipe" and "Temp" have any
#To study these impacts, a model cake.test1 is fitted. There appears to be no substantial recipe impact
##The sequential ANOVA output (anova(cake.test1)) assists in assessing the influence of each component
#Cake.test2, a second model, is fitted by reversing the order of "Temp" and "Recipe." ANOVA is run on t
```

Question 2f

Two variance tables are constructed, both of which produce symmetrical outcomes. These findings indicate that "Temp" appears to have a considerable influence on its own. When "Recipe" is considered (adjusted for "Recipe"), the relevance of "Temp" lessens.

In contrast, "Recipe" stays important even after adjusting for "Temp." In conclusion, the research shows that "Temp" alone looks important, but its relevance lowers when "Recipe" is considered. However, even when "Temp" is taken into consideration, "Recipe" remains significant.

"'