STAT6180: Assignment (Semester 2)

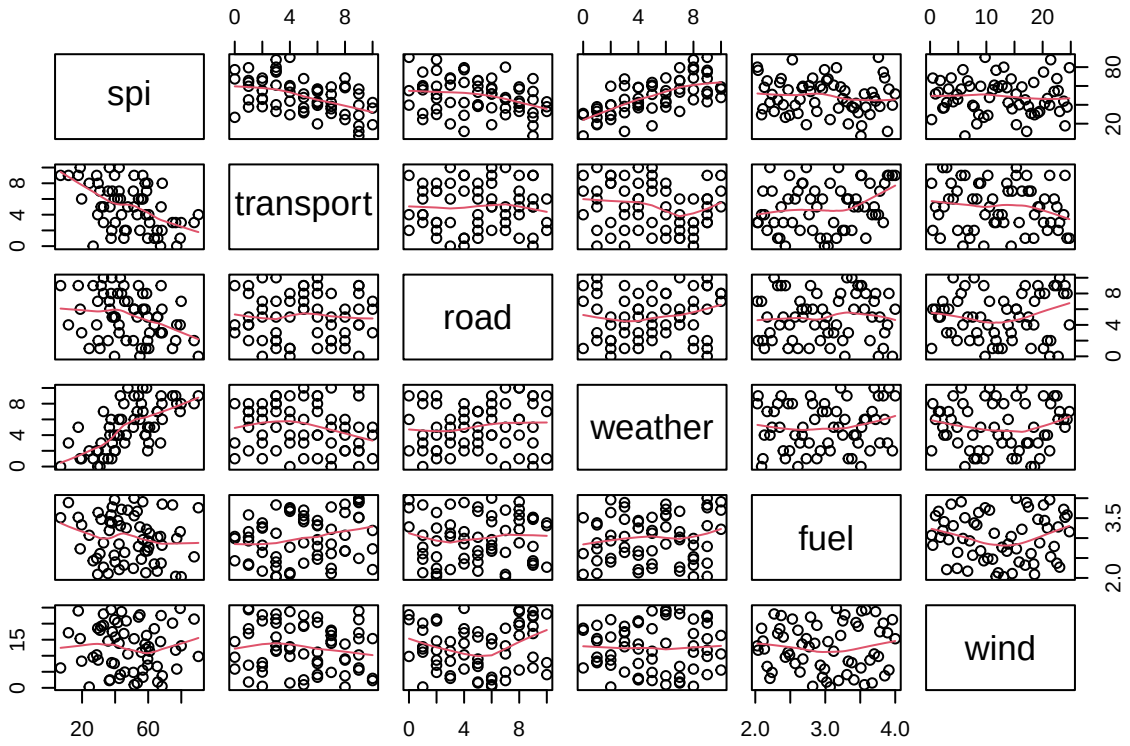
Josiah Jackson

2023-10-09

Question 1

a) Produce a plot and a correlation matrix of the data. Comment on possible relationships between the response and predictors and relationships between the predictors themselves

```
# Import data and produce plot
traffic <- read.csv("data/traffic.csv", header = TRUE)
pairs(traffic, panel = panel.smooth)
```



Regression Model:

$$\hat{spi} = 62.8071 - 2.175 \times transport - 2.4097 \times road + 4.256 \times weather - 3.6145 \times fuel - 0.1358 \times wind$$

Comments*:* We can observe a high correlation between spi and the variables transport and weather

```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
```

Comments*:* The correlation matrix shows high positive correlation between spi and weather, and moderate negative correlation with variables transport and road. Between the predictors, there is only a small positive correlation between transport and fuel.

b) Fit a model using all the predictors to explain the spi response. Then, using the full model, estimate the impact of weather on spi. Do this by producing a 95% confidence interval that quantifies the change in spi for every one index value increase of weather and comment.

```r
# Fit model using all predictors
spi.lm <- lm(spi ~ ., data = traffic)
summary(spi.lm)
```

```
##
## 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
# Confidence interval to estimate the impact of weather on spi
b_humidity <- spi.lm$coefficients[4]
n <- nrow(traffic)
tquant <- qt(0.975, n - 2)
se_b_humidity <- summary(spi.lm)$coefficients[,2][4]

lower <- b_humidity - tquant * se_b_humidity
upper <- b_humidity + tquant * se_b_humidity

paste(lower, upper)
```

```
## [1] "3.35096433921561 5.14032296064003"
```

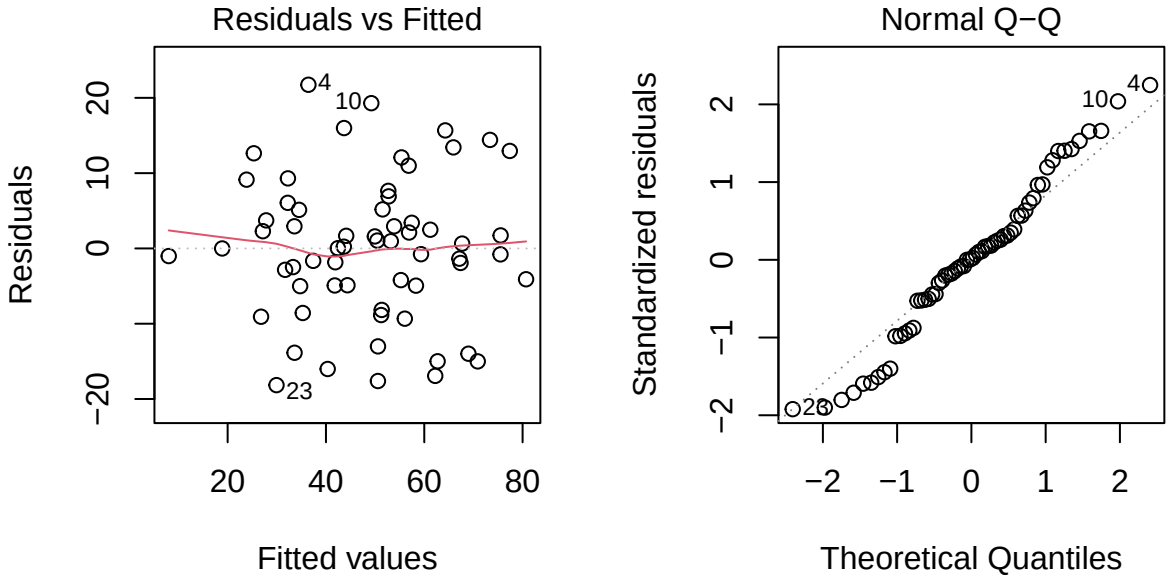
Confidence Interval:

$$\beta weather \pm t \times s.e.(\beta weather) = 4.2456 \pm 2.0003 \times 0.4473 = (3.35096433921561, 5.14032296064003)$$

Comments*:* For each index value that weather increases, it is expected that the spi index value will increase between 3.351 and 5.140 values.

c) Conduct an F-test for the overall regression (i.e. is there any relationship between the response and the predictors)

```r
# Check assumptions
par(mfrow = c(1, 2))
plot(spi.lm, which = 1:2)
```



Comments*:* The Residuals vs Fitted plot contains no discernible pattern, and the Normal Q-Q plot demonstrates a visible linear trend, indicating that the residuals are close to normally distributed. Constant variance and normality assumptions are therefore satisfied.

Multiple Regression Model:

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

Parameters:

$\hat{spi}$ is the dependent variable

$\beta_0$is the intercept

$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are representative of the coefficients of the independent variables

Independent variables: transport, road, weather, fuel, wind

ε : residuals/error term

Hypotheses:

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_1 : \text{not all } \beta_i \text{ are equal}$$

```
# ANOVA table
spi.aov <- anova(spi.lm)
spi.aov
```

```
# Create reduced overall ANOVA table
reg_SS <- sum(spi.aov$`Sum Sq`)
reg_df <- 5
reg_MS <- reg_SS / reg_df
res_SS <- spi.aov$`Sum Sq`[6]
res_df <- 56
res_MS <- res_SS / res_df
Fobs <- reg_MS / res_MS
Pval <- pf(Fobs, reg_df, res_df, lower.tail = FALSE)

Pr_F <- 0
if (Pval < 0.000000001) {
  Pr_F = 0
} else {
  Pr_F = Pval
}

reduced_aov <- matrix(c(reg_df, reg_SS, reg_MS, Fobs, Pr_F,
                        res_df, res_SS, res_MS, NaN, NaN), ncol=5, byrow=TRUE)
colnames(reduced_aov) <- c('Df', 'Sum Sq', 'Mean Sq', 'F Value', 'Pr(>F)')
rownames(reduced_aov) <- c('Regression', 'Residuals')

reduced_aov
```

```
##            Df    Sum Sq    Mean Sq  F Value Pr(>F)
## Regression  5 21206.153 4241.23065 43.16294      0
## Residuals  56  5502.613   98.26094      NaN    NaN
```

$$\text{Test Statistic} : F_{obs} = \frac{MS_{Reg}}{MS_{Res}} = \frac{4241.231}{98.261} = 43.163$$

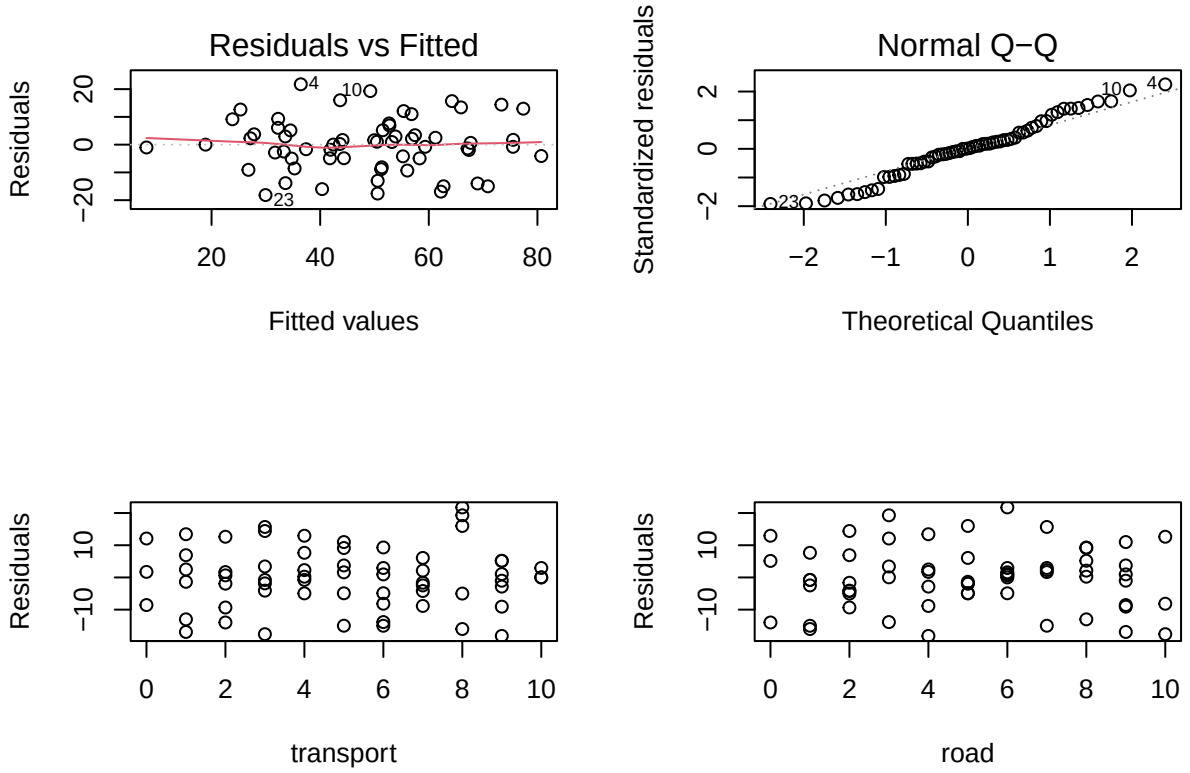
Null Distribution: The null distribution for the test statistic is $F_{5,56}$

P-Value : $P(F_{5,56} = 43.16294) = 0 = 5.256094 \times 10^{-}18 < 0.05$

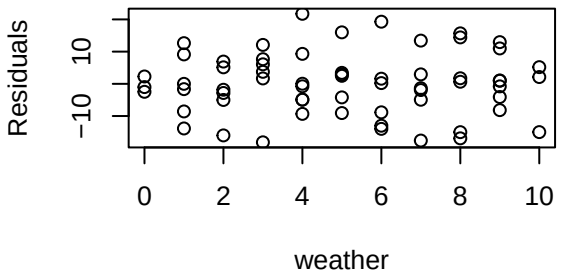
Conclusion*:* Since the P-Value is significantly smaller than the level of significance, there is enough evidence to reject H_0. This means that there is a significant linear relationship between spi and at least one of the five predictor variables

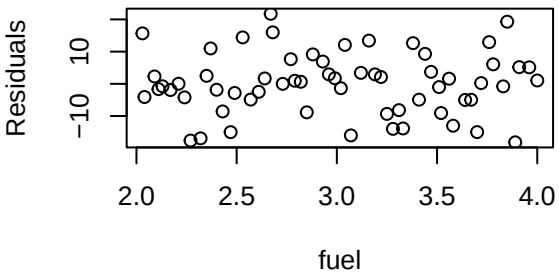
d) Validate the full model and comment on whether the full regression model is appropriate to explain the spi

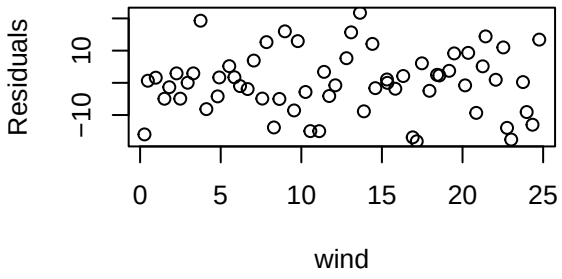
```
par(mfrow = c(2, 2))
plot(spi.lm, which = 1:2)
plot(resid(spi.lm) ~ transport, data = traffic, xlab = "transport", ylab = "Residuals")
plot(resid(spi.lm) ~ road, data = traffic, xlab = "road", ylab = "Residuals")
```




```
par(mfrow = c(2, 2))
plot(resid(spi.lm) ~ weather, data = traffic, xlab = "weather", ylab = "Residuals")
plot(resid(spi.lm) ~ fuel, data = traffic, xlab = "fuel", ylab = "Residuals")
plot(resid(spi.lm) ~ wind, data = traffic, xlab = "wind", ylab = "Residuals")
```








```
print(spi.aov)
```

```
## 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
```

Comments: Both the fuel and wind predictor variables both have P-Values of over 0.05, and are therefore insignificant. This means that it is not appropriate to use the full model to explain the spi. A new regression model without the fuel and wind variables will be used to proceed

```
spi.lm2 <- lm(spi ~ transport + road + weather, data = traffic)
spi.aov2 <- anova(spi.lm2)
spi.aov2
```

Comments: After removing the fuel and wind predictor variables, all the predictors in the new model are significant. This model is more appropriate to explain the spi.

e) Find the R^2 and comment on what it means in the context of this dataset

```
summary(spi.lm)
```

```
##
## 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
```

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{15703.5}{21206.1} = 0.7405$$

Comments: This significantly high R-squared value shows that the predictor variables in the original model contribute significantly to the spi. It means that 74.05% of the variation in the data set is explained by the full linear regression model.

f) Using model selection procedures discussed in the unit, find the best multiple regression model that explains the data. State the final fitted regression model.

```
summary(spi.lm)$coefficients
```

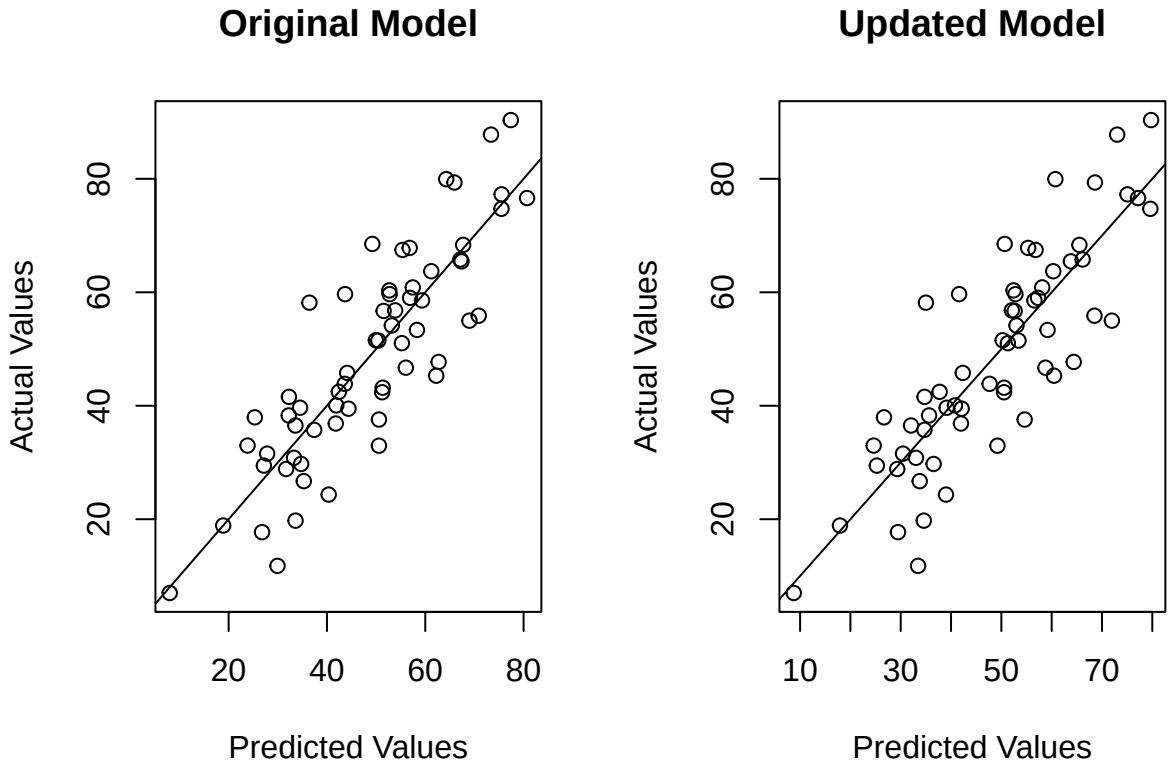
```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 62.8071258  7.4079572  8.4783327 1.270384e-11
## transport   -2.1750456  0.4611007 -4.7170731 1.634444e-05
## road        -2.4096862  0.4365099 -5.5203467 9.036105e-07
## weather      4.2456436  0.4472731  9.4922858 2.918068e-13
## fuel        -3.6145270  2.2759093 -1.5881682 1.178790e-01
## wind        -0.1357686  0.1764484 -0.7694523 4.448584e-01
```

```r
summary(spi.lm2)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 51.737015  4.1026793 12.610543 2.909752e-18
## transport   -2.321620  0.4449080 -5.218204 2.535898e-06
## road        -2.456284  0.4394267 -5.589748 6.399379e-07
## weather      4.144978  0.4463447  9.286496 4.480360e-13
```

```r
par(mfrow = c(1, 2))
plot(x=predict(spi.lm), y=traffic$spi,
     xlab='Predicted Values',
     ylab='Actual Values',
     main='Original Model')
abline(a=0, b=1)

plot(x=predict(spi.lm2), y=traffic$spi,
     xlab='Predicted Values',
     ylab='Actual Values',
     main='Updated Model')
abline(a=0, b=1)
```



Comments: by comparing the original multiple regression model with the updated one, it is clear that the current updated one is the best model to explain the data. This is because all the predictor variables P-values are significantly smaller than 0.05. The Predicted vs Actual Value plots also show a line of better fit for the updated model.

g) Comment on the R^2 and adjusted R^2 in the full and final model you chose in part f. In particular explain why those goodness of fitness measures change

```
summary(spi.lm2)
```

```
##
## 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
```

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{15387.2}{21206.1} = 0.7405$$

Comments: This significantly high R-squared value shows that the predictor variables in the new model (transport, road and weather) contribute significantly to the spi. The adjusted R-squared value is extremely similar (0.7114) compared to the original model's (0.7174), meaning it is not too evident from these values which model is more reliable. However, since higher R-squared values are not always indicative of a better model, and the prediction vs actual value graphs show a better line of fit on the newer model, it is safe to assume that this model is more reliable than the original

Question 2

a) For this study, is the design balanced or unbalanced?

```
# Import data and check number of replicates across all the levels factors
cake <- read.csv("data/cake.csv", header = TRUE, stringsAsFactors=TRUE)
table(cake[, c("Temp", "Recipe")])
```

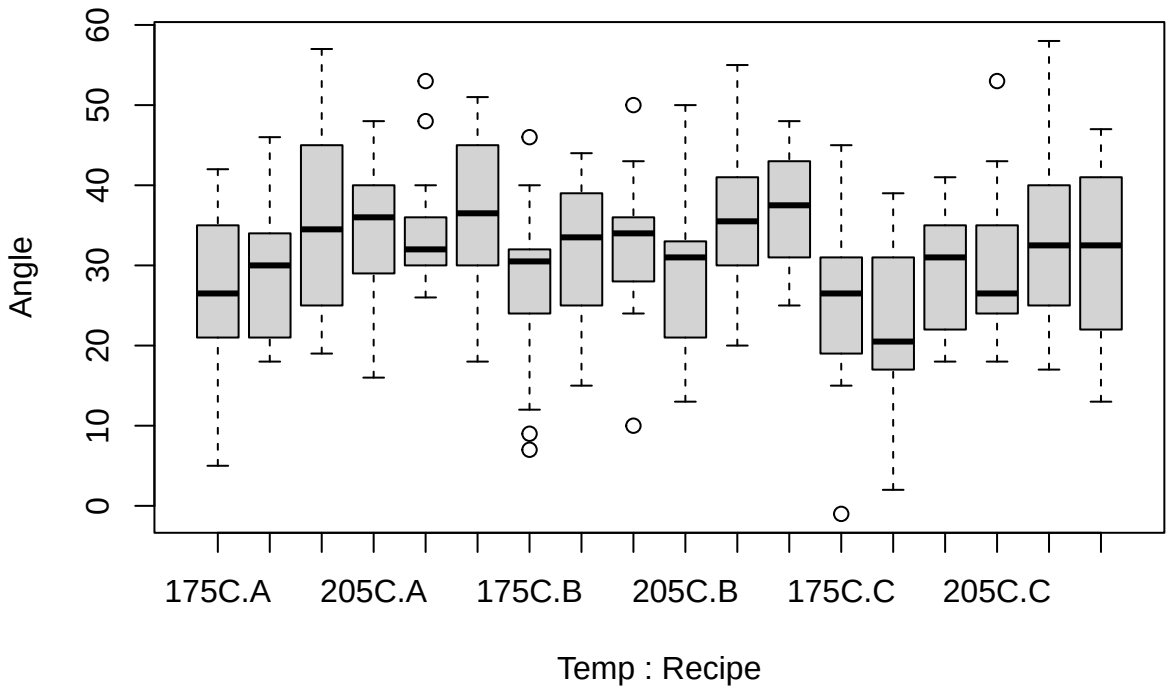
```
##       Recipe
## Temp    A  B  C
##   175C 14 14 14
##   185C 14 14 14
##   195C 14 14 14
```

```
##    205C 14 14 14
##    215C 14 14 14
##    225C 14 14 14
```

Comments: From the above results, we can see that design is balanced as it has an equal number of replicates for each combination of levels of the two factors

b) Construct two different preliminary graphs that investigate different features of the data and comment

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



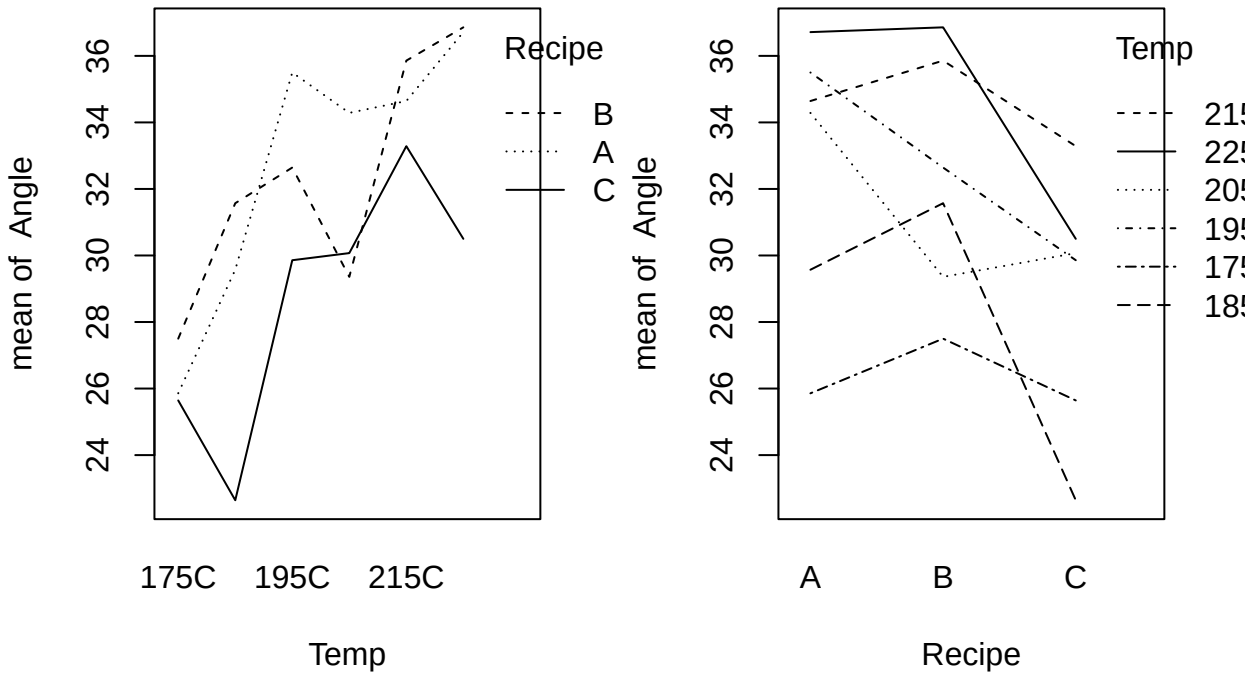
Comments: The boxplot shows mostly somewhat equal variance among levels, but there are some potential outliers. Will need to calculate the standard deviation for each level to compare.

```
# Check
tempVals <- unique(cake$Temp)
recipeVals <- unique(cake$Recipe)
for (temp in tempVals) {
  for (recipe in recipeVals) {
    col <- cake[cake$Temp == temp & cake$Recipe == recipe,]
    print(sd(col$Angle))
  }
}
```

```
## [1] 9.035948
## [1] 8.6799
## [1] 10.3968
## [1] 11.66756
## [1] 11.46734
## [1] 11.97181
## [1] 11.79798
## [1] 9.459282
## [1] 7.304973
## [1] 8.55236
## [1] 10.44846
## [1] 9.722807
## [1] 7.722167
## [1] 8.493048
## [1] 11.78936
## [1] 10.42503
## [1] 6.948792
## [1] 10.73993
```

Comments: the largest standard deviation is less than twice the smallest standard deviation, so the equal variance assumption is valid.

```
# Interaction plots
par(mfrow = c(1, 2))
with(cake, interaction.plot(Temp, Recipe, Angle))
with(cake, interaction.plot(Recipe, Temp, Angle))
```



Comments: both interaction plots show non-parallel lines for the means of each group at different levels of the independent variables, which indicates an interaction effect between the two independent variables.

c) Write down the full mathematical model for this situation, defining all appropriate parameters

The full Two-Way ANOVA model with interaction is:

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

with the parameters as:

$$Y_{ijk} : \text{the angle at which the cake broke}$$

$$\alpha_i : \text{The Recipe effect, there are two levels - A, B, C}$$

$$\beta_j : \text{The Temp effect, there are 6 levels - 175C, 185C, 195C, 205C, 215C, 225C}$$

$$\gamma_{ij} : \text{interaction effect between Recipe and Temp}$$

$$\epsilon_{ijk} : \text{the unexplained variation}$$

d) Analyse the data to study the effect of Temp and Recipe on breaking Angle of cake at 5% significance level

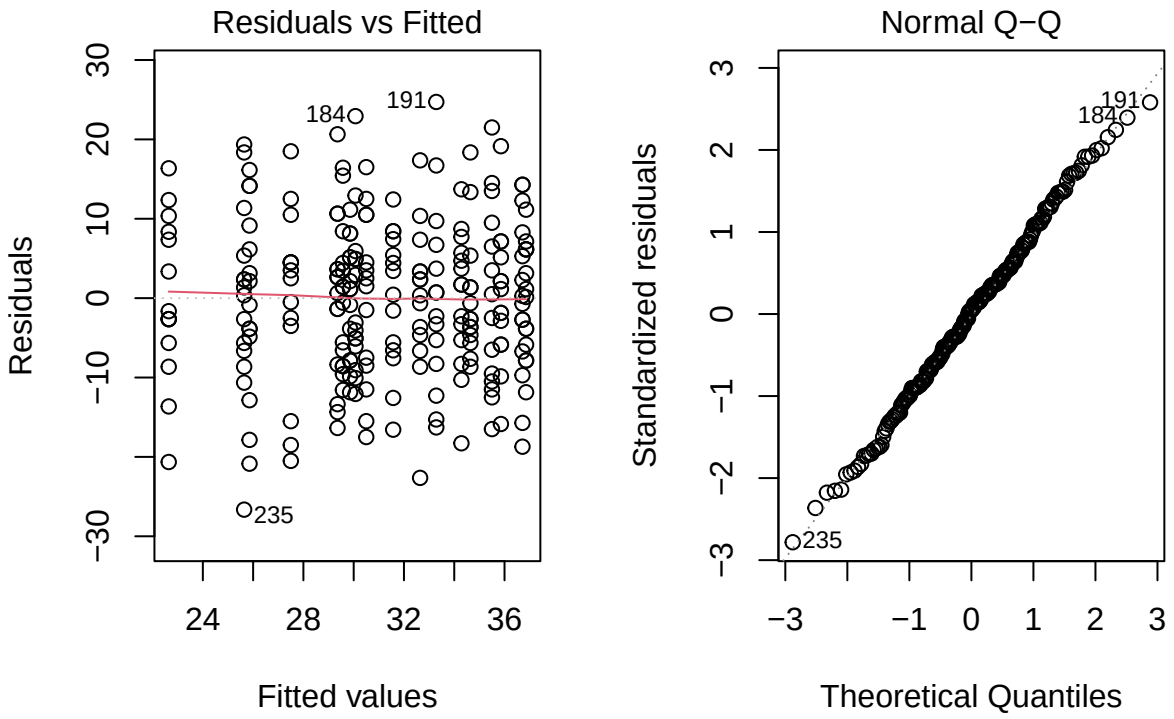
Hypotheses:

$$H_0 : \gamma_{ij} = 0 \text{for all } i, j$$

$$H_1 : \text{at least one } \gamma_{ij} \neq 0$$

```
# Fit the interaction model
cake.int <- lm(Angle ~ Recipe * Temp, data = cake)

# Validate interaction model with diagnostic plots
par(mfrow = c(1, 2))
plot(cake.int, which = 1:2)
```



Comments: the residuals are close to linear in the Normal QQ plot, so the normality assumption should be valid. The residual plot shows an equal spread so the constant variance assumption should also be valid.

```
# Run two-way ANOVA
print(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
```

```
summary(cake.int)
```

```
##
## Call:
## lm(formula = Angle ~ Recipe * Temp, data = cake)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -26.6429  -6.5714   0.4643   6.1429  24.7143
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       25.8571     2.6562   9.735  < 2e-16 ***
## RecipeB            1.6429     3.7565   0.437  0.66227
## RecipeC           -0.2143     3.7565  -0.057  0.95456
## Temp185C           3.7143     3.7565   0.989  0.32380
## Temp195C           9.6429     3.7565   2.567  0.01088 *
## Temp205C           8.4286     3.7565   2.244  0.02578 *
## Temp215C           8.7857     3.7565   2.339  0.02019 *
## Temp225C          10.8571     3.7565   2.890  0.00421 **
## RecipeB:Temp185C   0.3571     5.3124   0.067  0.94646
## RecipeC:Temp185C  -6.7143     5.3124  -1.264  0.20753
## RecipeB:Temp195C  -4.5000     5.3124  -0.847  0.39782
## RecipeC:Temp195C  -5.4286     5.3124  -1.022  0.30790
## RecipeB:Temp205C  -6.5714     5.3124  -1.237  0.21733
## RecipeC:Temp205C  -4.0000     5.3124  -0.753  0.45224
## RecipeB:Temp215C  -0.4286     5.3124  -0.081  0.93577
## RecipeC:Temp215C  -1.1429     5.3124  -0.215  0.82985
## RecipeB:Temp225C  -1.5000     5.3124  -0.282  0.77792
## RecipeC:Temp225C  -6.0000     5.3124  -1.129  0.25987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.939 on 234 degrees of freedom
## Multiple R-squared:  0.1479, Adjusted R-squared:  0.08595
## F-statistic: 2.388 on 17 and 234 DF,  p-value: 0.002021
```

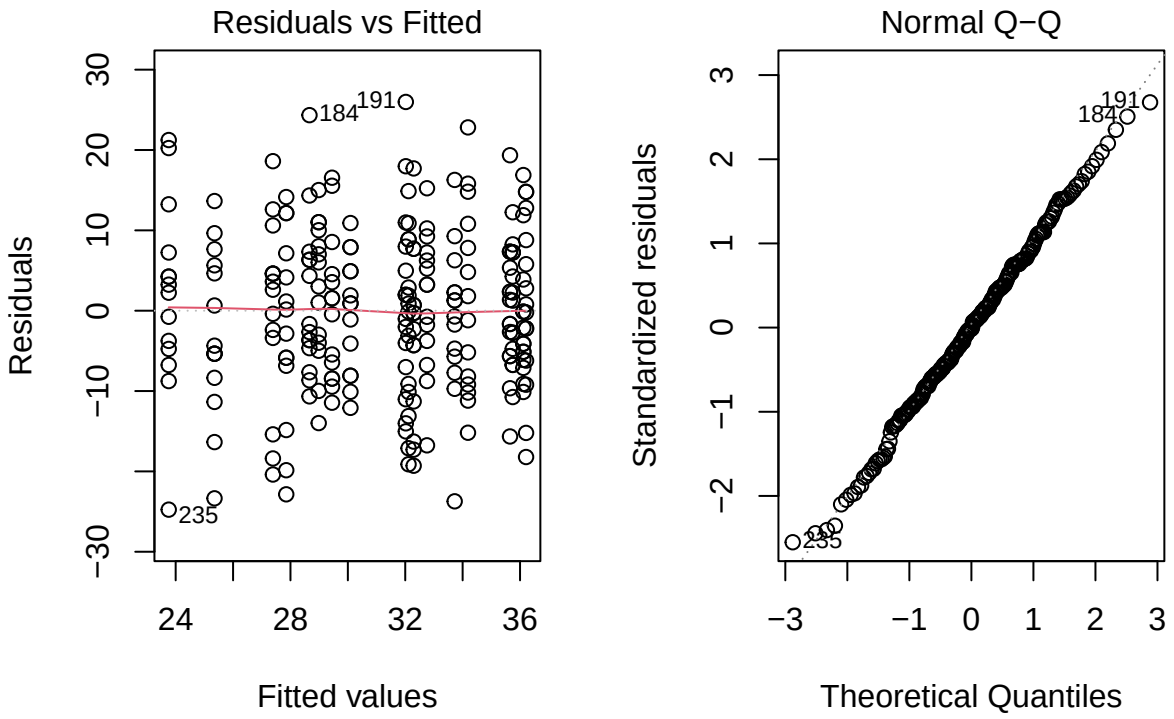
Comments: we can see that the interaction terms are insignificant since the F-test of the interaction term has a P-Value of 0.776 (3 d.p.). This means they can be removed from the model and we have not yet reached our final model.

e) Repeat the above test analysis for the main effects

```r
# Update model to only use main effects
cake.int2 = update(cake.int, . ~ . - Recipe:Temp)
summary(cake.int2)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 27.8531746   1.757827 15.8452319 2.320714e-39
## RecipeB     -0.4642857   1.522323 -0.3049851 7.606375e-01
## RecipeC     -4.0952381   1.522323 -2.6901248 7.635780e-03
## Temp185C     1.5952381   2.152889  0.7409754 4.594209e-01
## Temp195C     6.3333333   2.152889  2.9417829 3.577337e-03
## Temp205C     4.9047619   2.152889  2.2782228 2.357852e-02
## Temp215C     8.2619048   2.152889  3.8375889 1.583160e-04
## Temp225C     8.3571429   2.152889  3.8818263 1.334932e-04
```

```r
# Validate interaction model with diagnostic plots
par(mfrow = c(1, 2))
plot(cake.int2, which = 1:2)
```



Comments: the residuals are close to linear in the Normal QQ plot, so the normality assumption should be valid. The residual plot shows an equal spread so the constant variance assumption should also be valid.

```
# Run Two-Way ANOVA
print(anova(cake.int2))
```

```
## 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
```

Comments: we can see that the main effects are significant as they have respective P-values of 0.0140 and 0.0001 (4 d.p.). This means that they cannot be removed and that we have reached our final model.

f) State your conclusions about the effect of Temp and Recipe on the Angle response

Overall, the effect of the *recipe* of the cake on the *angle* at which the cake breaks does not depend on the *temperature* at which the cake was baked. Neither does the effect of the *temperature* at which the cake was baked depend on the *recipe* of the cake. However, both the *temperature* and the *recipe* variables do have a significant effect on the *angle* at which the cake breaks.

Also uploaded to:

https://github.com/MQ-STAT2170-6180-Assignment-S2-2023/assignment-s2-2023-jijackson111/blob/main/Assignment-45948763.pdf