MAT325 Project 4

Jordan Badstuebner

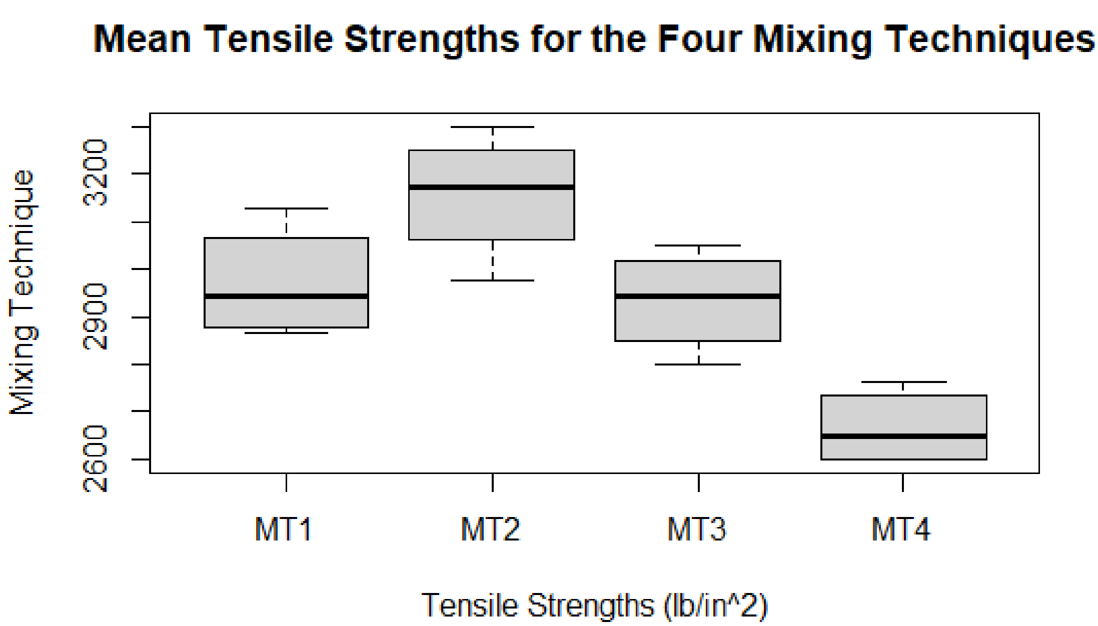
18 August 2020

```
#############
####  1  ####
#############
```

(a) Test the hypothesis that mixing techniques affect the strength of the cement. Use $\alpha = 0.05$.

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## df1$variable  3 489740  163247   12.73 0.000489 ***
## Residuals    12 153908   12826                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Construct a graphical display (boxplot) to compare the mean tinsile strengths for the four mixing techniques. What are your conclusions?

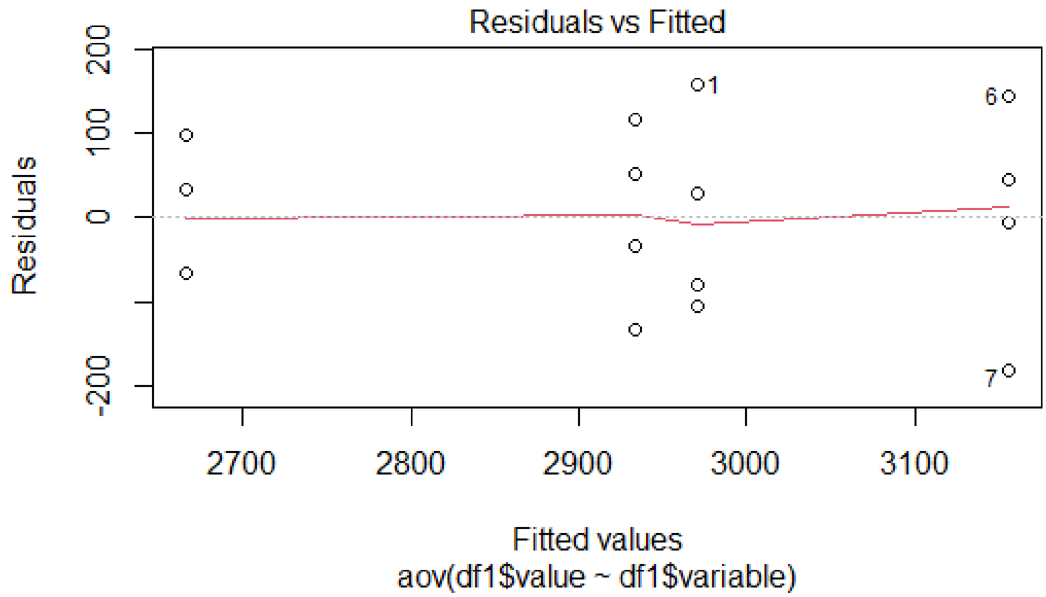


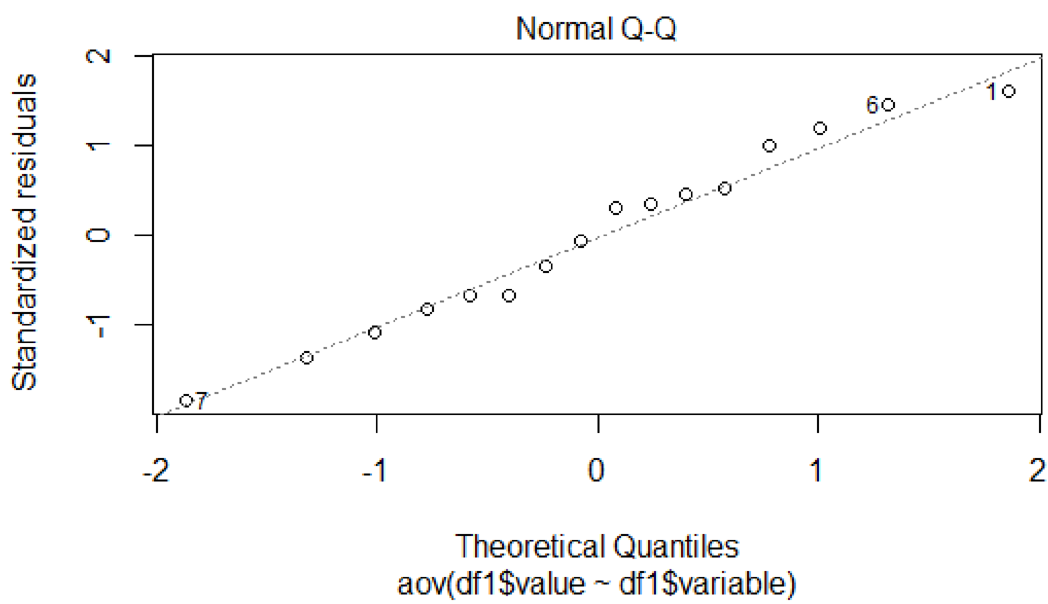
Based on the plot, we conclude that μ_1 and μ_3 are very close or the same; that μ_4 differs from μ_1 and μ_3, that μ_2 differs from μ_1 and μ_3, and that μ_2 and μ_4 are different.

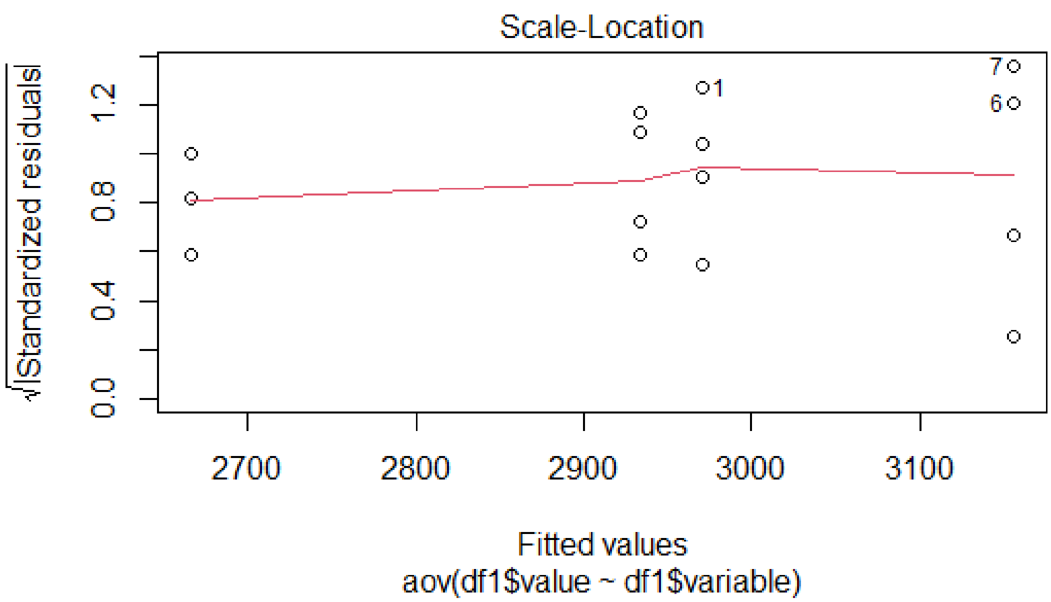
(c) Use the Fisher LSD method with $\alpha = 0.05$ to make comparisons between pairs of means.

```
## $statistics
##   MSerror Df     Mean       CV  t.value     LSD
##     12826 12 2931.812 3.862864 2.178813 174.482
##
## $parameters
##         test p.ajusted       name.t ntr alpha
##   Fisher-LSD      none df1$variable   4  0.05
##
## $means
##     df1$value       std r      LCL      UCL  Min  Max     Q25    Q50     Q
75
## MT1   2971.00 120.55704 4 2847.623 3094.377 2865 3129 2883.75 2945.0 3032.
25
## MT2   3156.25 135.97641 4 3032.873 3279.627 2975 3300 3106.25 3175.0 3225.
00
## MT3   2933.75 108.27242 4 2810.373 3057.127 2800 3050 2875.00 2942.5 3001.
25
## MT4   2666.25  80.97067 4 2542.873 2789.627 2600 2765 2600.00 2650.0 2716.
25
##
## $comparison
## NULL
##
## $groups
##     df1$value groups
## MT2   3156.25      a
## MT1   2971.00      b
## MT3   2933.75      b
## MT4   2666.25      c
##
## attr(,"class")
## [1] "group"
```

(d,e) Construct a normal probability plot of the residuals. What conclusion would you draw about the validity of the normality assumption? Plot the residuals versus the predicted tensile strength. Comment on the plot.

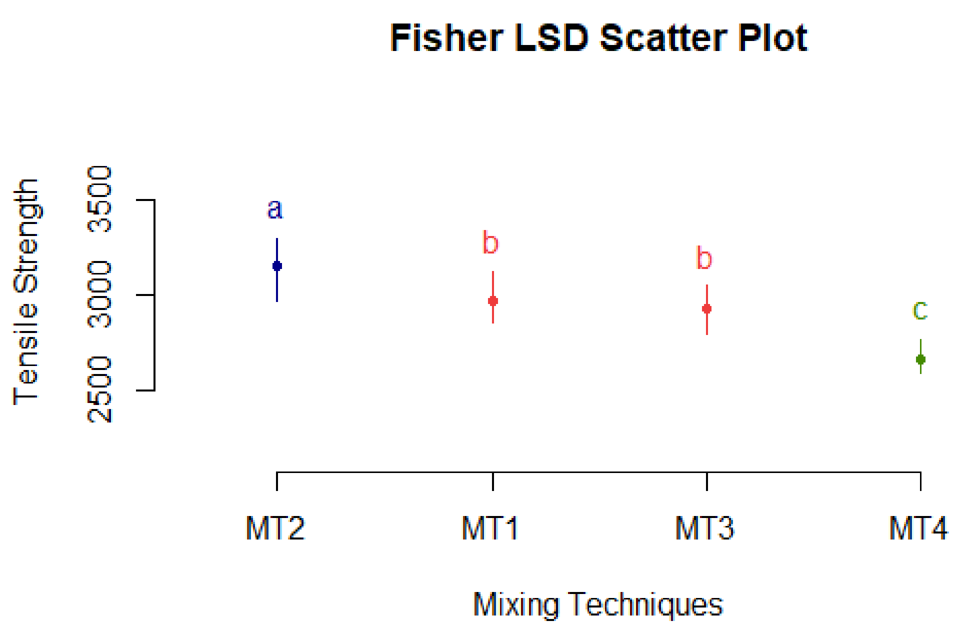

Residuals vs Fitted
Residuals
200
100
0
-200
2700
2800
2900
3000
3100
Fitted values
aov(df1$value ~ df1$variable)
1
6
7


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-1
0
1
2
Theoretical Quantiles
aov(df1$value ~ df1$variable)
6
1
7



There is nothing unusual about the normal Q-Q probability plot of residuals. Similarly, there is nothing unusual about the Residuals vs. Fitted (Predicted) or Scale-Location plots. All assumptions appear to be met.

(f) Prepare a scatter plot of the results to aid the interpretation of the results of this experiment.



Observe in the above scatter plot that our results from part(c) correspond to our results described in part(b).

(g) Rework part (c) using Tukey's test with $\alpha = 0.05$. Do you get the same conclusions from Tukey's test that you did from the graphical procedure and/or the Fisher LSD method?

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = df1$value ~ df1$variable)
##
## $`df1$variable`
##             diff        lwr        upr     p adj
## MT2-MT1  185.25  -52.50029  423.00029 0.1493561
## MT3-MT1  -37.25 -275.00029  200.50029 0.9652776
## MT4-MT1 -304.75 -542.50029  -66.99971 0.0115923
## MT3-MT2 -222.50 -460.25029   15.25029 0.0693027
## MT4-MT2 -490.00 -727.75029 -252.24971 0.0002622
## MT4-MT3 -267.50 -505.25029  -29.74971 0.0261838
```

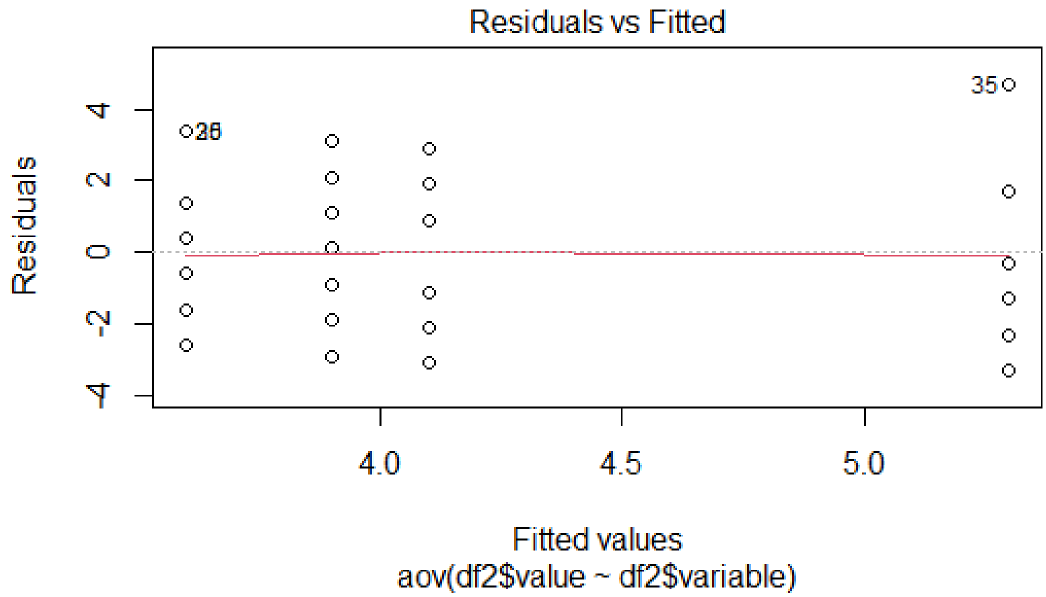
From the pairwise comparison, it is clear that the pair "Mixing Technique 3 vs. Mixing Technique 1" is not significantly different. All the other pairs of means are significantly different.

```
#############
####  2  ####
#############
```

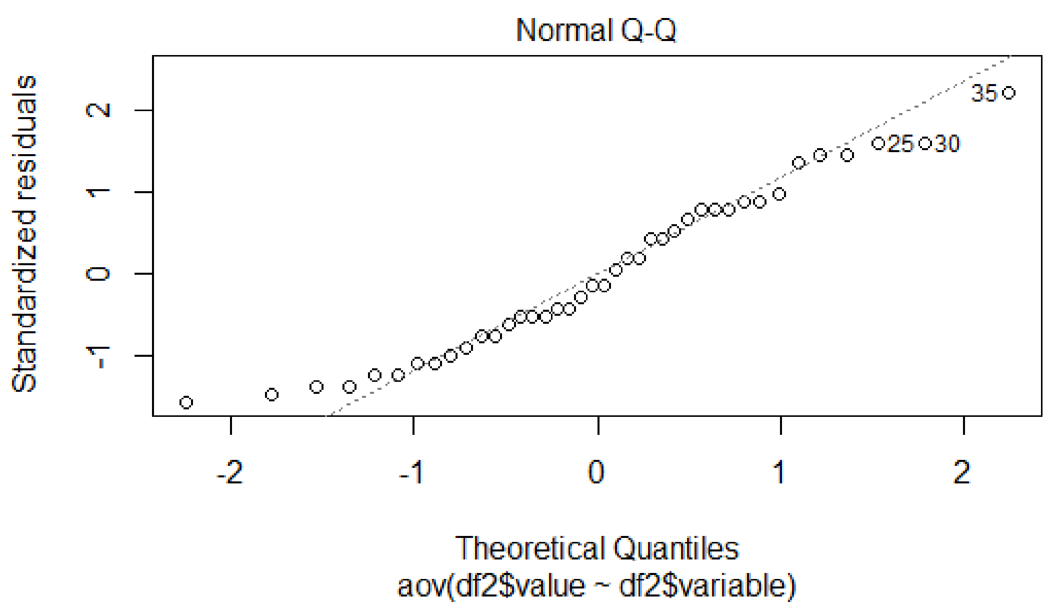
(a)) Is there evidence to support a claim that the type of car rented affects the length of the rental contract? Use $\alpha = 0.05$. If so, which types of cars are responsible for the difference?

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## df2$variable  3  16.67   5.558    1.11  0.358
## Residuals    36 180.30   5.008
```

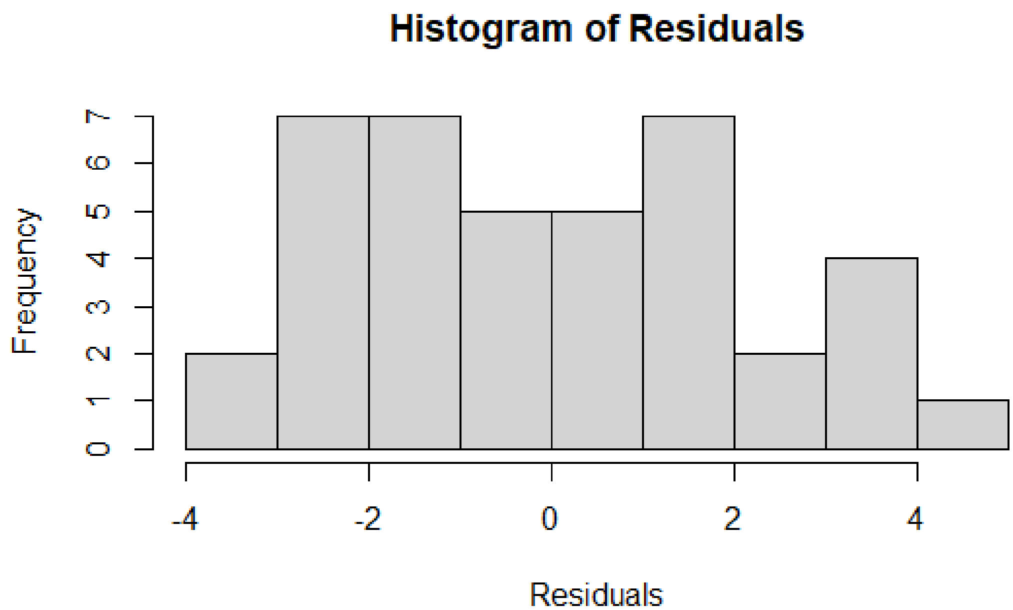
(b) Analyze the residuals from this experiment and comment on the model adequacy.



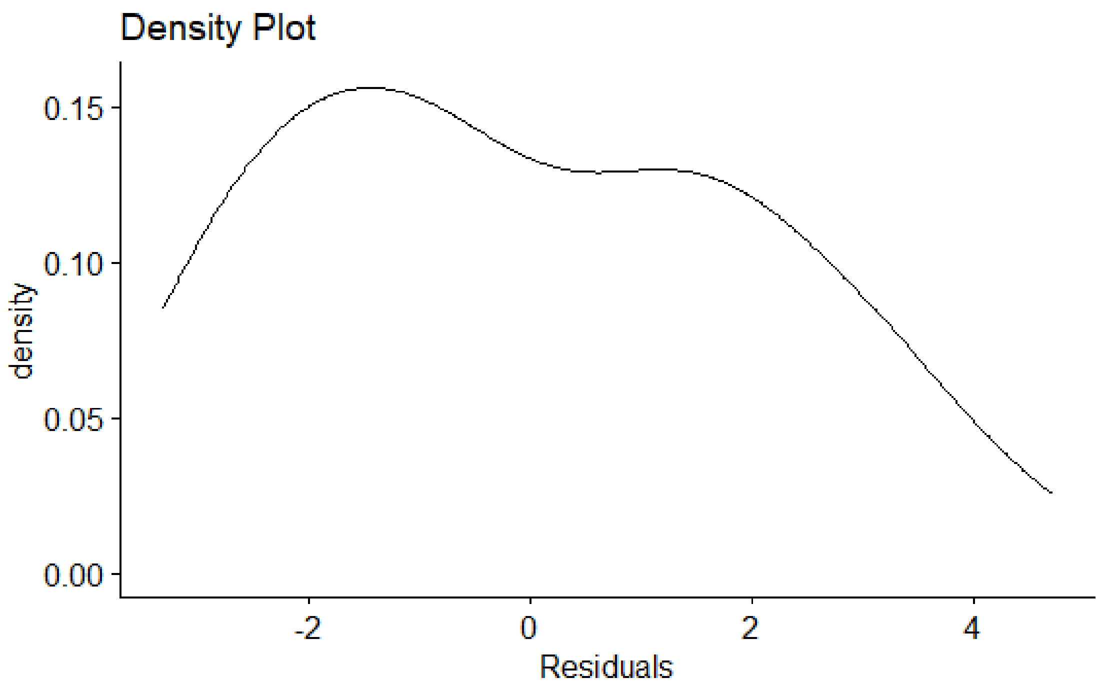
When we look at the Residuals vs. Fitted plot, the residuals appear to be mostly evenly distributed. The red smooth line is horizontal. This suggests normality and homogeneity of variance, however points 26 and 35 are detected as outliers, which may be affecting the results.



The residuals in the Normal Q-Q plot appear reasonably tight to the reference line, but do take on a slight “heavy tails” shape which could signify that the data has more extreme values than what would be expected if the residuals were truly normal distributed.



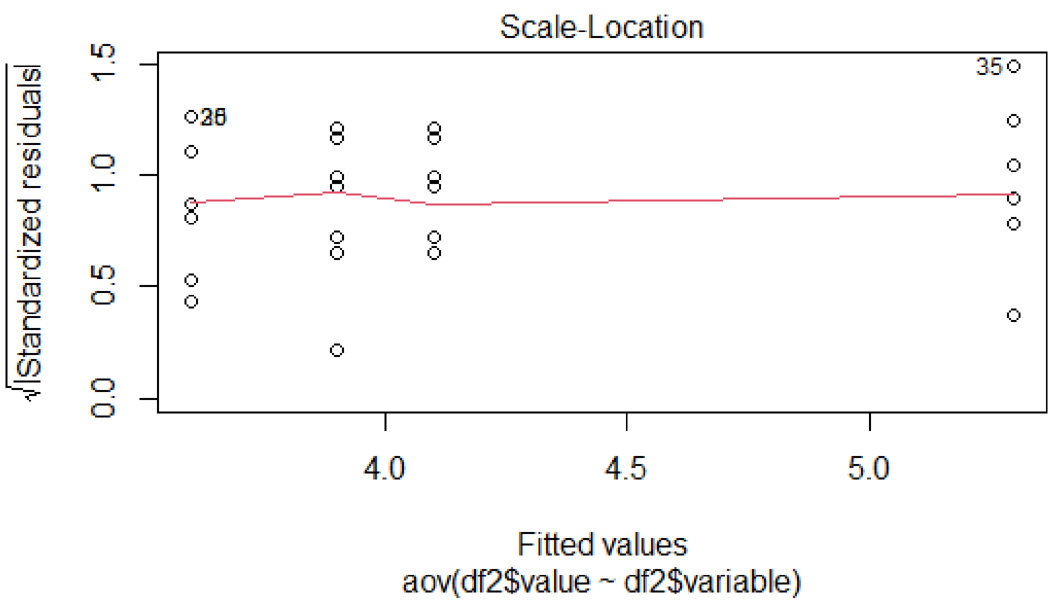
The Histogram of Residuals plot seems to loosely fit a “bell curve” shape, althought the dip in the center could be cause for concern. If the residuals are not normally distributed, it means the data is not normally distributed, which is contradictory to the assumption of the model. Let’s do some more in depth analysis:



The Density plot also appears to be bell curved. To be certain, let's use the Shapiro-Wilk method for normality testing.

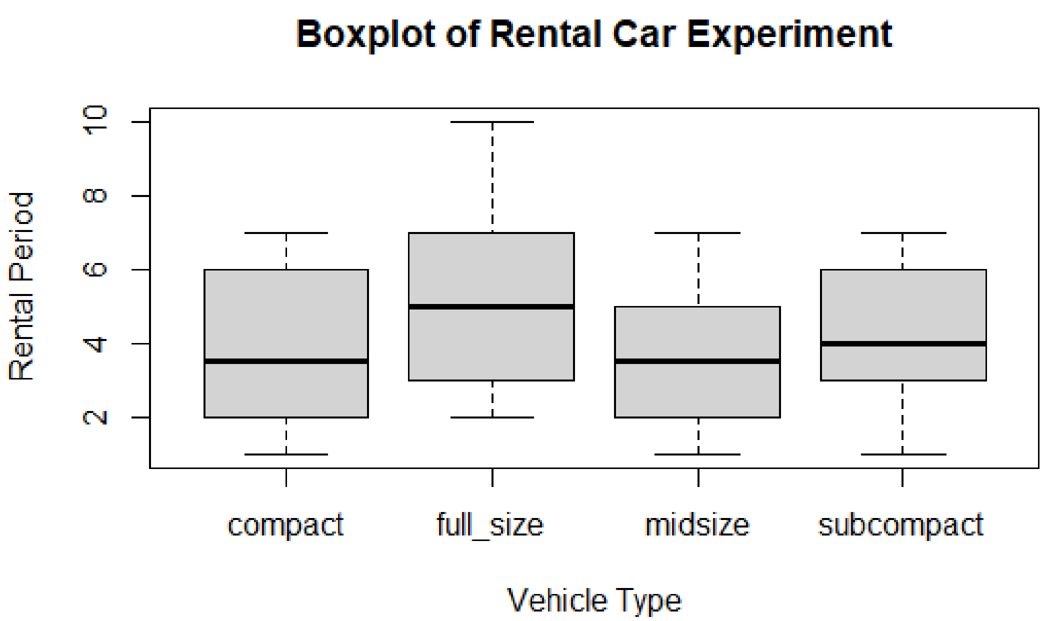
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(res.aov_2)
## W = 0.95761, p-value = 0.1386
```

From the output, $p - val = 0.1386 > 0.05$ implies that the distribution of the data are not significantly different from the normal distribution and we can therefore assume normality of the residuals.



In the Scale-Location plot, the residuals appear to be mostly evenly distributed. The red smooth line is horizontal. It appears the residuals are spead equally along the ranges of predictors and therefore meet the assumption of homoscedasticity.

(c) Which type of car have the longest rental period?



Given the boxplot above, full size vehicles clearly have the longest mean rental period.

```
#############
####   3   ####
#############
```

(a) Analyze the data from this experiment.

```
##
## Call:
## lm(formula = df3$value ~ df3$variable + df3$additional_factor,
##     data = df3)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.039867 -0.008967  0.003133  0.010667  0.022333
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.496867   0.015694  31.660 1.08e-09 ***
## df3$variableoil_2       0.048600   0.014530   3.345 0.010157 *
## df3$variableoil_3       0.008800   0.014530   0.606 0.561529
## df3$additional_factor2  0.118667   0.018758   6.326 0.000226 ***
## df3$additional_factor3 -0.017667   0.018758  -0.942 0.373847
## df3$additional_factor4 -0.128000   0.018758  -6.824 0.000135 ***
## df3$additional_factor5  0.004667   0.018758   0.249 0.809795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02297 on 8 degrees of freedom
## Multiple R-squared:  0.959,  Adjusted R-squared:  0.9283
## F-statistic:  31.2 on 6 and 8 DF,  p-value: 3.958e-05
```

The P value of the overall model is <0.0001. So the model is significant with an $R^2 = 0.96$, explaining almost 96% variation.

```
## Analysis of Variance Table
##
## Response: df3$value
##                       Df   Sum Sq   Mean Sq F value    Pr(>F)
## df3$variable           2 0.006706 0.0033529  6.3527   0.02229 *
## df3$additional_factor  4 0.092100 0.0230249 43.6257 1.781e-05 ***
## Residuals              8 0.004222 0.0005278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

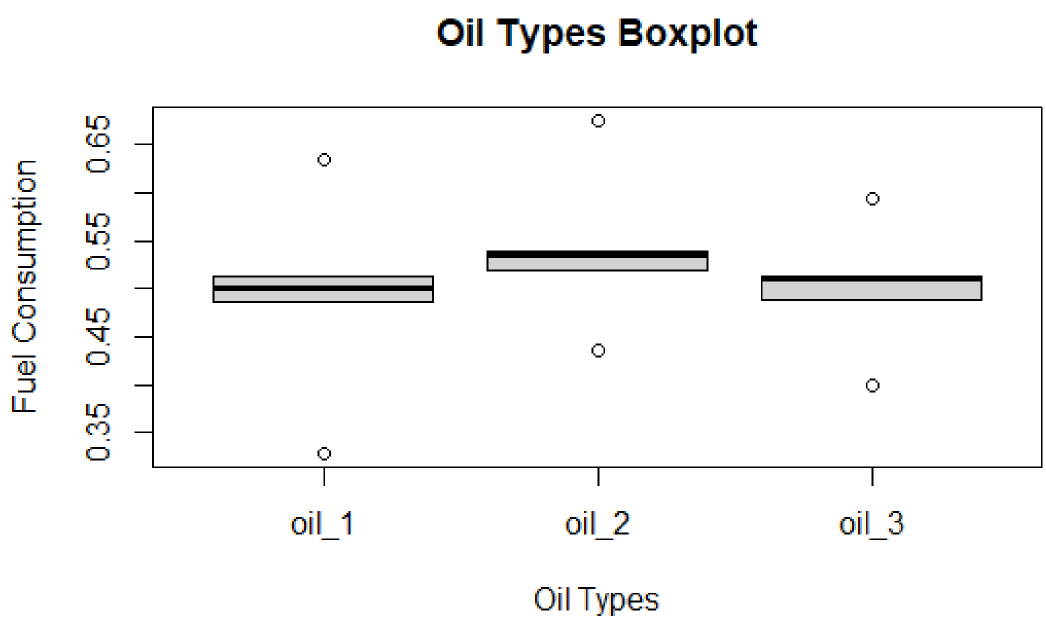
The $\{P - val\}_{oil} = 0.02229 < \alpha = 0.05$, so we can conclude that the oils differ significantly in terms of fuel-efficiency.

(b) Use the Bonferroni method to make comparisons among the three lubricating oils to determine specifically which oils differ in break-specific fuel consumption.

```
##
##  Pairwise comparisons using t tests with pooled SD
```

```
##
## data:  df3$value and df3$variable
##
##       oil_1 oil_2
## oil_2 1     -
## oil_3 1     1
##
## P value adjustment method: bonferroni
```

The Bonferroni Method seems to suggest that any oil type is not statistically superior to any other, which is contradictory to our ANOVA results. Let's take a closer look

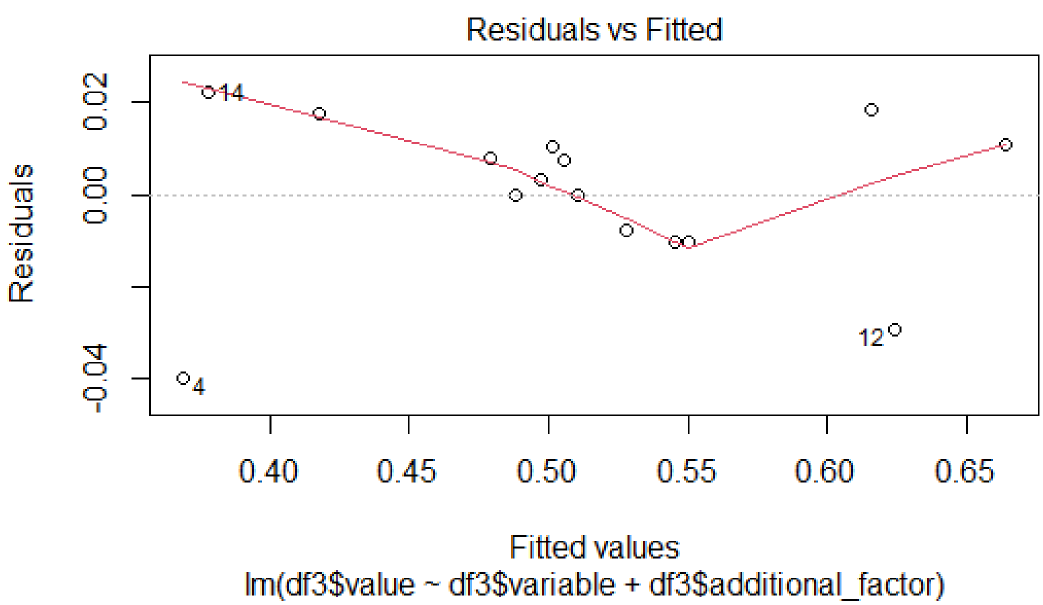



```
## $statistics
##     MSerror Df      Mean       CV  t.value        LSD
##   0.0005278  8 0.5115333 4.491183 2.306004 0.03350617
##
## $parameters
##         test p.ajusted       name.t ntr alpha
##   Fisher-LSD      none df3$variable   3  0.05
##
## $means
##       df3$value        std r       LCL       UCL   Min   Max   Q25   Q50
Q75
## oil_1    0.4924 0.10865220 5 0.4687076 0.5160924 0.329 0.634 0.487 0.500 0
.512
## oil_2    0.5410 0.08612491 5 0.5173076 0.5646924 0.435 0.675 0.520 0.535 0
.540
## oil_3    0.5012 0.06969720 5 0.4775076 0.5248924 0.400 0.595 0.488 0.510 0
.513
```

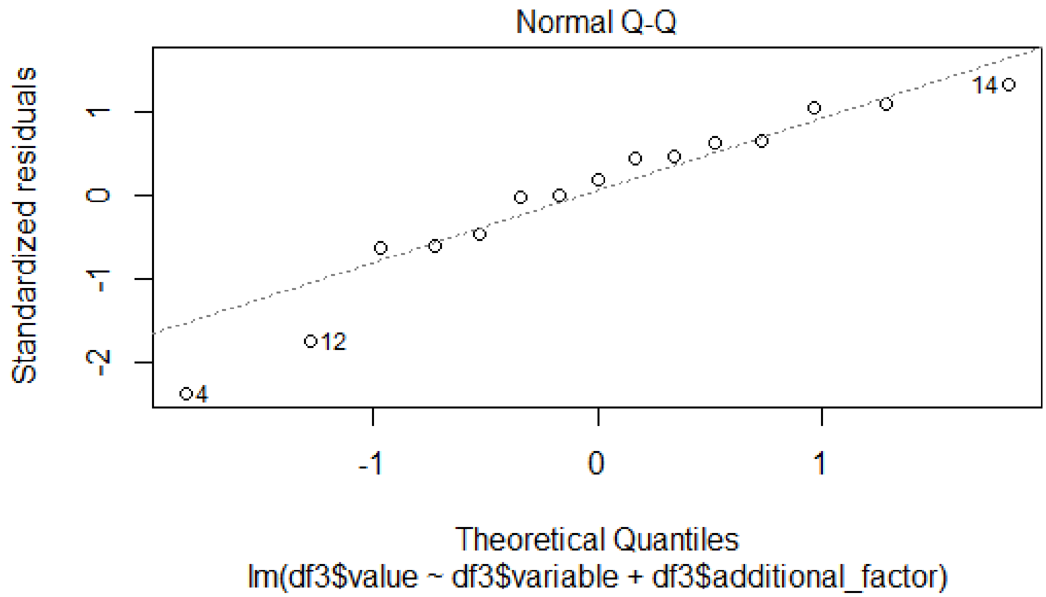
```
##
## $comparison
## NULL
##
## $groups
##       df3$value groups
## oil_2    0.5410      a
## oil_3    0.5012      b
## oil_1    0.4924      b
##
## attr(,"class")
## [1] "group"
```

The Boxplot and Fisher LSD results both suggest that Oil 2 is different from oils 1 and 3, while oils 1 and 3 do not differ from one another. The boxplot does reveal some outliers which may be skewing that data. I would opt to investigate these outliers prior to making the claim that oil 2 is the superior oil.

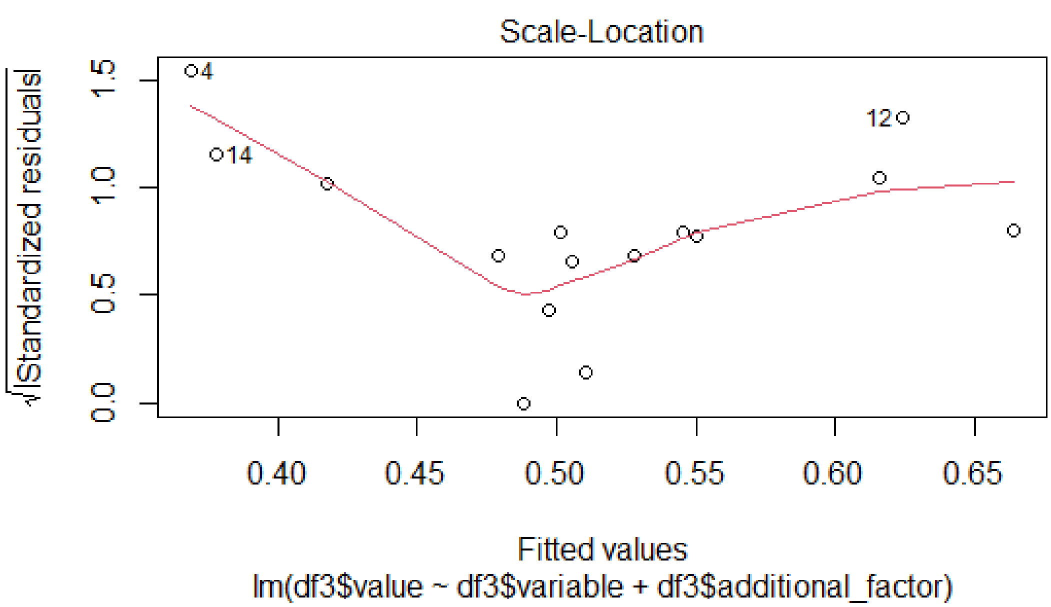
(c) Analyze the residuals from this experiment.



When we look at the Residuals vs. Fitted plot, the data appears to be clustered about the center. Points 14, 4, and 12 are detected as outliers and may be severely affecting the normality and homogeneity of variance.



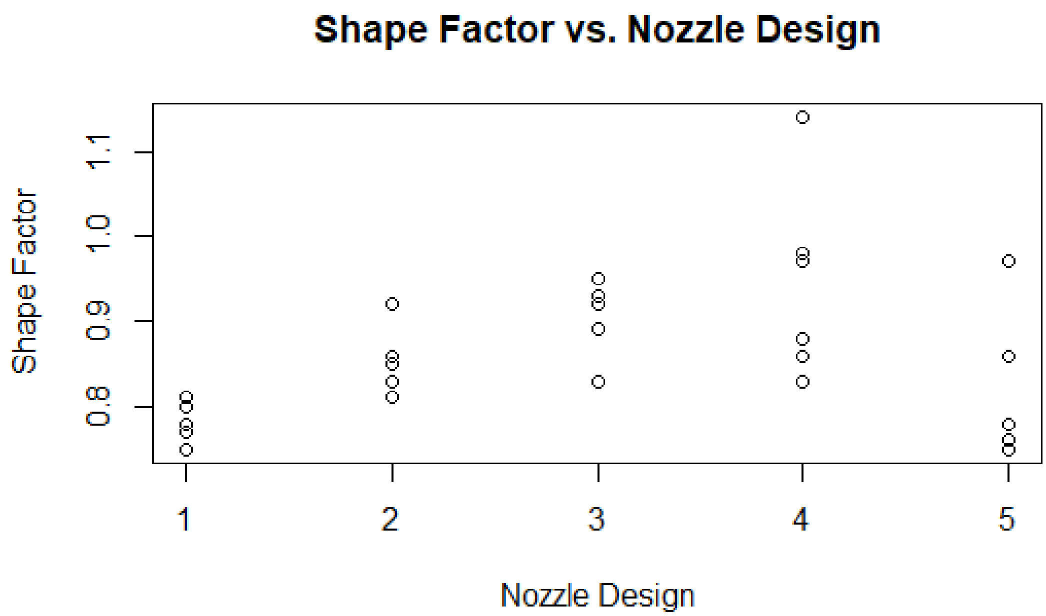
The Normal Q-Q plot shows good alignment. This is a good indication that the residuals are normally distributed.



In the Scale-Location plot, the residuals appear wide outside and condensed in the middle. This is creating a v-shaped red smooth line. It does not appear the residuals are spead equally along the ranges of predictors and therefore does not appear to meet the assumption of homoscedasticity.

```
#############
####  4  ####
#############
```

(a) Does nozzle design affect the shape factor? Compare nozzles with a scatter plot and with an analysis of variance, using α = 0.05.



The scatterplot above suggests that nozzle design does indeed affect shape factor.

```
##
## Call:
## lm(formula = df4$value ~ df4$variable + df4$additional_factor,
##     data = df4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.078667 -0.024167 -0.001833  0.028083  0.121333
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.85700    0.03090  27.732  < 2e-16 ***
## df4$variable2                0.07167    0.03090   2.319 0.031091 *
## df4$variable3                0.12000    0.03090   3.883 0.000925 ***
## df4$variable4                0.16167    0.03090   5.231 4.05e-05 ***
## df4$variable5                0.03167    0.03090   1.025 0.317736
## df4$additional_factor14.37 -0.05400    0.03385  -1.595 0.126360
## df4$additional_factor16.59 -0.04600    0.03385  -1.359 0.189329
## df4$additional_factor20.43 -0.10600    0.03385  -3.131 0.005259 **
## df4$additional_factor23.46 -0.11600    0.03385  -3.427 0.002672 **
```

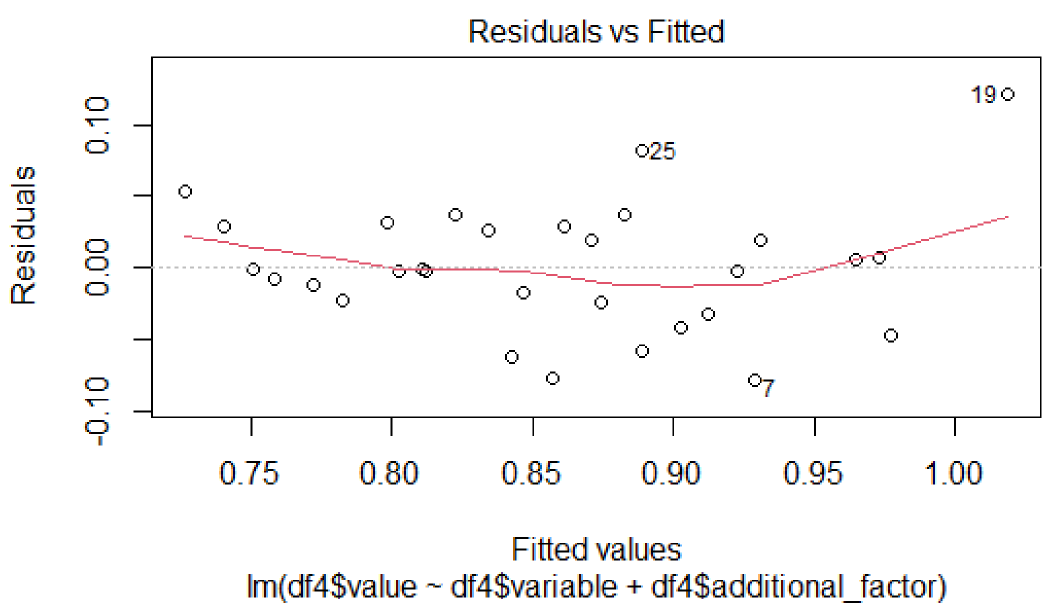
```
## df4$additional_factor28.74 -0.13000    0.03385  -3.840 0.001022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05353 on 20 degrees of freedom
## Multiple R-squared:  0.7423, Adjusted R-squared:  0.6263
## F-statistic: 6.401 on 9 and 20 DF,  p-value: 0.0002787
```

The P value of the overall model is 0.0002787, so the model is significant though an $R^2 = 0.7423$ only explains approximately 74% variation.

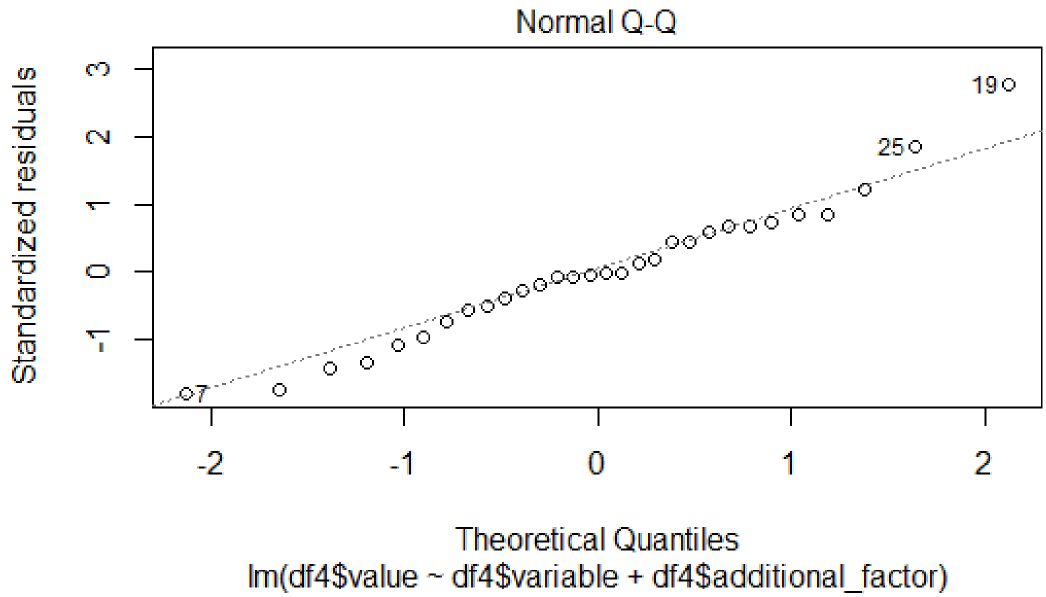
```
## Analysis of Variance Table
##
## Response: df4$value
##                         Df   Sum Sq  Mean Sq F value    Pr(>F)
## df4$variable             4 0.102180 0.025545  8.9162 0.0002655 ***
## df4$additional_factor    5 0.062867 0.012573  4.3886 0.0073642 **
## Residuals               20 0.057300 0.002865
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Nozzle Type $P - val = 0.0002655 < \alpha = 0.05$, so we can conclude that the nozzle types differ significantly in terms of shape factor.

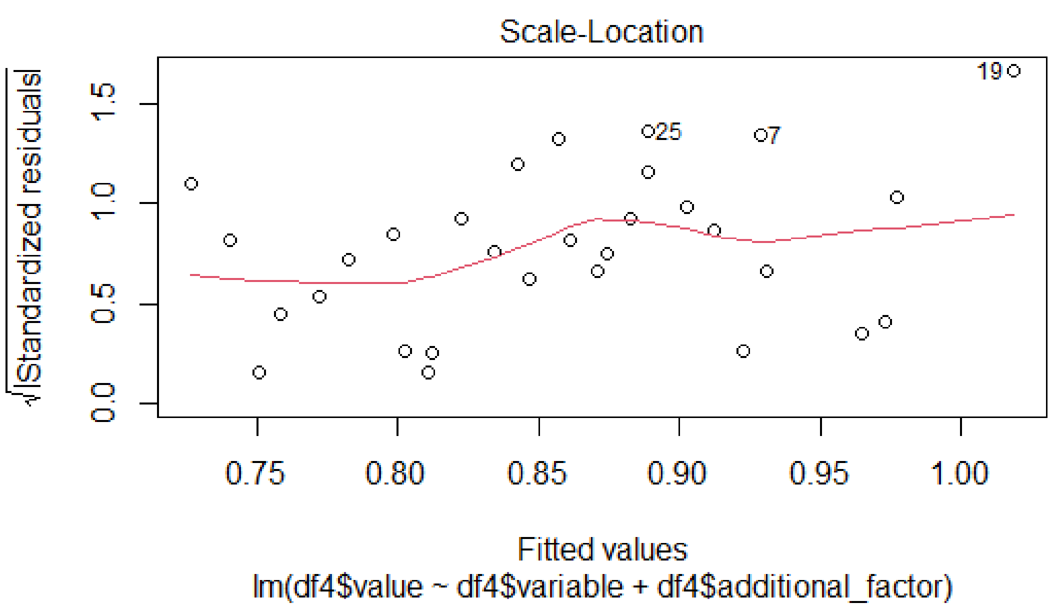
(b) Analyze the residual from this experiment.



When we look at the Residuals vs. Fitted plot, the residuals appear to be mostly evenly distributed. The red smooth line is close to horizontal. This suggests normality and homogeneity of variance, however points 25 and 19 are detected as outliers, which may be affecting the results.



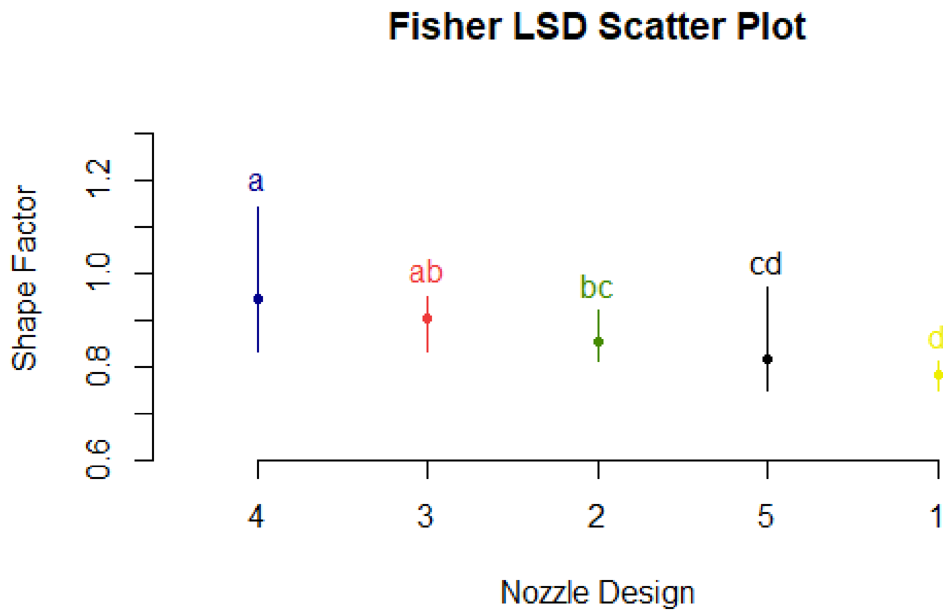
The Normal Q-Q plot shows good alignment with a slight veering in the lower left. This seems to indicate that the residuals are normally distributed.



In the Scale-Location plot, the residuals appear to be fairly evenly distributed with small clusters in the lower left and upper middle. The red smooth line is almost horizontal. It appears the residuals are spread equally along the ranges of predictors and seem to satisfy the assumption of homoscedasticity.

(c) Which nozzle designs are different with respect to shape factor?

```
## $statistics
##    MSerror Df      Mean       CV  t.value        LSD
##   0.002865 20 0.8586667 6.233582 2.085963 0.06446268
##
## $parameters
##         test p.ajusted       name.t ntr alpha
##   Fisher-LSD      none df4$variable   5  0.05
##
## $means
##   df4$value        std r       LCL       UCL  Min  Max    Q25   Q50    Q75
## 1 0.7816667 0.02136976 6 0.7360847 0.8272487 0.75 0.81 0.7725 0.780 0.7950
## 2 0.8533333 0.03723797 6 0.8077513 0.8989153 0.81 0.92 0.8350 0.850 0.8575
## 3 0.9016667 0.04215052 6 0.8560847 0.9472487 0.83 0.95 0.8900 0.905 0.9275
## 4 0.9433333 0.11360751 6 0.8977513 0.9889153 0.83 1.14 0.8650 0.925 0.9775
## 5 0.8133333 0.08664102 6 0.7677513 0.8589153 0.75 0.97 0.7600 0.770 0.8400
##
## $comparison
## NULL
##
## $groups
##   df4$value groups
## 4 0.9433333      a
## 3 0.9016667     ab
## 2 0.8533333     bc
## 5 0.8133333     cd
## 1 0.7816667      d
##
## attr(,"class")
## [1] "group"
```



The differences in nozzle types with respect to shape factor are shown above in the Fisher LSD Scatter plot.