Ch-03 R Codes

Ping-Yang Chen

Textbook: Montgomery, D. C. (2012). *Design and analysis of experiments*, 8th Edition. John Wiley & Sons.

Online handouts: https://github.com/PingYangChen/ANOVA_Course_R_Code

Chapter 3

One-way ANOVA

Read the csv file `3_PlasmaEtching.csv` in R. Make sure that in the `data.frame` the variable `Power` is a factor. If not sure, apply `as.factor()` function to set the property of the variable `Power` after reading the dataset.

```r
df1 <- read.csv(file.path("data", "3_PlasmaEtching.csv"))
df1$Power <- as.factor(df1$Power)
```

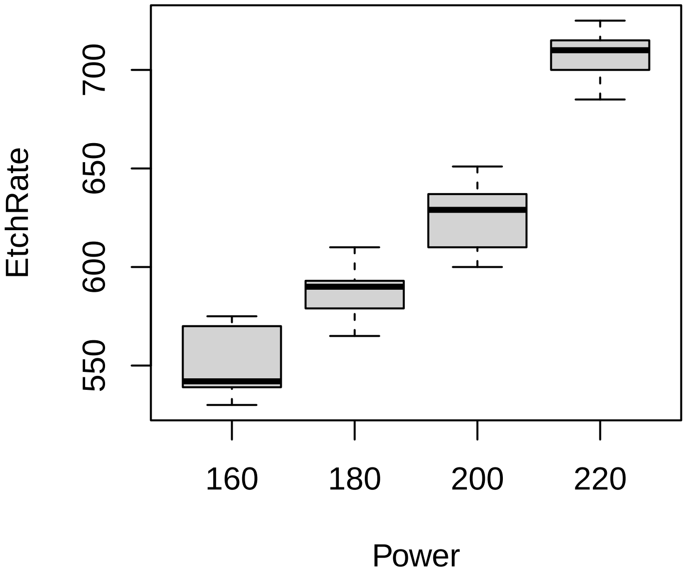
To compute descriptive statistics of the data in each subgroup of a dataset in R, we use `tapply()`.

```r
tapply(df1$EtchRate, df1$Power, summary)
```

```
## $`160`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   530.0   539.0   542.0   551.2   570.0   575.0 
## 
## $`180`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   565.0   579.0   590.0   587.4   593.0   610.0 
## 
## $`200`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   600.0   610.0   629.0   625.4   637.0   651.0 
## 
## $`220`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     685     700     710     707     715     725
```

Alternatively, boxplots provide a quick and direct means of observing the differences among the responses of the four treatments (groups or levels of a factor).

```r
# Draw the grouped boxplot
boxplot(EtchRate ~ Power, data = df1)
```



The function `aov()` fits the ANOVA model. For one-way ANOVA, the command is as follows. Then, we call `summary()` to examine the ANOVA table.

```
fit <- aov(EtchRate ~ Power, data = df1)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Power        3  66871   22290    66.8 2.88e-09 ***
## Residuals   16   5339     334                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Adequacy Checking

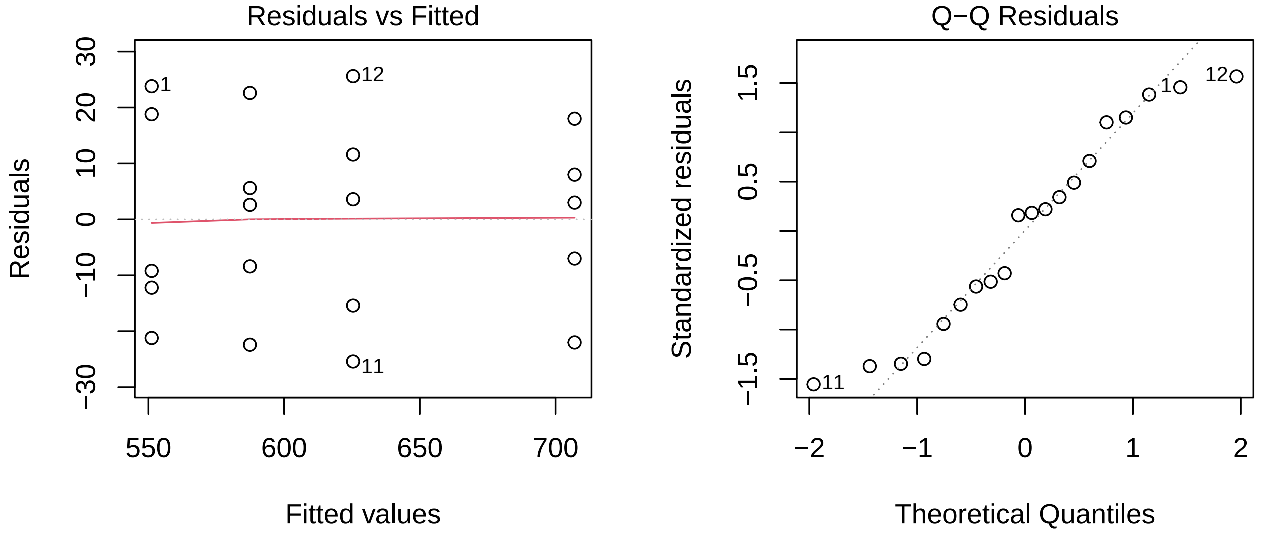
The adequacy of an ANOVA model can be studied from residual plots. The basic approach is to use the `plot()` function with the fitted ANOVA model object as its input argument. Since there are four residual plots, we can use `par(mfrow = c(2, 2))` before the `plot()` function to view all of them simultaneously.

The first (upper left) plot is the residual plot against the fitted values. This plot is used to check the consistency of the variance with changes in the fitted value. A lack of any visually obvious pattern in the dots on the plot is desired.

The second (upper right) plot is the residuals' Normal Quantile-Quantile (QQ) plot. Ideally, the dots form a straight line.

The remaining two plots at the bottom are standardized residuals against the fitted values and standardized residuals against the factor levels, respectively. They are also used to check the consistency of the variance.

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



Post-ANOVA Comparison of Means

The estimate of the overall mean μ and the Power's treatment effects τ_1 to τ_4 are

$$\hat{\mu} = \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} = \bar{y}_{..};$$

$$\hat{\tau}_1 = \frac{1}{n}\sum_{j=1}^{n} y_{1j} - \hat{\mu} = \bar{y}_{1.} - \bar{y}_{..}; \hat{\tau}_2 = \frac{1}{n}\sum_{j=1}^{n} y_{2j} - \hat{\mu} = \bar{y}_{2.} - \bar{y}_{..};$$

$$\hat{\tau}_3 = \frac{1}{n}\sum_{j=1}^{n} y_{3j} - \hat{\mu} = \bar{y}_{3.} - \bar{y}_{..}; \hat{\tau}_4 = \frac{1}{n}\sum_{j=1}^{n} y_{4j} - \hat{\mu} = \bar{y}_{4.} - \bar{y}_{..}.$$

The R codes are as follows.

```
mean(df1$EtchRate) # Overall
mean(df1$EtchRate[df1$Power == 160]) - mean(df1$EtchRate) # tau_1
mean(df1$EtchRate[df1$Power == 180]) - mean(df1$EtchRate) # tau_2
mean(df1$EtchRate[df1$Power == 200]) - mean(df1$EtchRate) # tau_3
mean(df1$EtchRate[df1$Power == 220]) - mean(df1$EtchRate) # tau_4
```

Following an ANOVA in which we have rejected the null hypothesis of equal treatment means, we wish to test all pairwise mean comparisons:

$$H_0: \; \mu_i = \mu_j$$
$$H_1: \; \mu_i \neq \mu_j$$

for all $i \neq j$. Here, we introduce three approaches.

Pairwise t-tests

The straightforward approach to test for all pairs of the hypotheses is to conduct the Pairwise t-tests simultaneously. The following codes give the results under Bonferroni adjustment on the p-value.

```
pairwise.t.test(df1$EtchRate, df1$Power, p.adjust = "bonferroni")
```

```
##
```

```
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df1$EtchRate and df1$Power
##
##     160     180     200
## 180 0.038   -       -
## 200 5.1e-05 0.028   -
## 220 2.2e-09 1.0e-07 1.6e-05
##
## P value adjustment method: bonferroni
```

Tukey's Test

Tukey's procedure makes use of the distribution of the studentized range statistic

$$q = \frac{\bar{y}_{max} - \bar{y}_{min}}{\sqrt{MS_E/n}}$$

where $\bar{y}_{max}$ and $\bar{y}_{min}$ are the largest and smallest sample means respectively, out of a group of p sample means. For equal sample sizes, Tukey's test declares two means significantly different if the absolute value of their sample differences exceeds

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}$$

where $q_\alpha(a, f)$ is the upper α percentage points of q and f is the number of degrees of freedom associated with the MS_E. For more insights on the distribution of q, please refer to the textbook. Tukey's method is performed by the function `TukeyHSD()`.

```r
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = EtchRate ~ Power, data = df1)
##
## $Power
##          diff        lwr       upr     p adj
## 180-160  36.2   3.145624  69.25438 0.0294279
## 200-160  74.2  41.145624 107.25438 0.0000455
## 220-160 155.8 122.745624 188.85438 0.0000000
## 200-180  38.0   4.945624  71.05438 0.0215995
## 220-180 119.6  86.545624 152.65438 0.0000001
## 220-200  81.6  48.545624 114.65438 0.0000146
```

Fisher's LSD Method

The R package **agricolae** provides the function `LSD.test()` to perform Fisher's LSD test. Adjustment for the P-value is necessary. Typically, we set `p.adj = "bonferroni"` for the Bonferroni method.

```r
if (!("agricolae" %in% rownames(installed.packages()))) {
  install.packages("agricolae")
}
library(agricolae)
out <- LSD.test(fit, "Power", p.adj = "bonferroni")
print(out)
```

```
## $statistics
##   MSerror Df   Mean       CV  t.value      MSD
##     333.7 16 617.75 2.957095 3.008334 34.75635
```

```
##
## $parameters
##         test  p.ajusted name.t ntr alpha
##   Fisher-LSD bonferroni  Power   4  0.05
##
## $means
##     EtchRate      std r       se      LCL      UCL Min Max Q25 Q50 Q75
## 160    551.2 20.01749 5 8.169455 533.8815 568.5185 530 575 539 542 570
## 180    587.4 16.74216 5 8.169455 570.0815 604.7185 565 610 579 590 593
## 200    625.4 20.52559 5 8.169455 608.0815 642.7185 600 651 610 629 637
## 220    707.0 15.24795 5 8.169455 689.6815 724.3185 685 725 700 710 715
##
## $comparison
## NULL
##
## $groups
##     EtchRate groups
## 220    707.0      a
## 200    625.4      b
## 180    587.4      c
## 160    551.2      d
##
## attr(,"class")
## [1] "group"
```

The most important parts of the outputs are shown below:

- `$means` displays the estimated mean of the etching rate at each level of power.
- `$groups` indicates the significance of the difference in the etching rate at each level of power. The column groups in `$groups` encodes the treatment levels with no significant difference in the etching rate by the same alphabet letter.

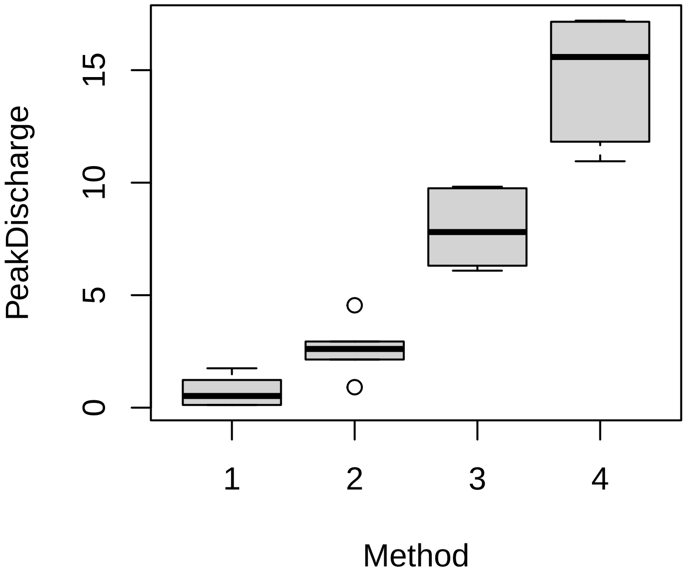
Variance-Stabilizing Transformations: Peak Discharge Experiments

Read the csv file `3_PeakDischarge.csv` in R. Make sure that in the `data.frame` the variable `Method` is a factor because it represents categorical data (i.e., different methods). If not sure, apply `as.factor()` function to set the property of the variable `Method` after reading the dataset.

```r
# Load the dataset from the 'data' directory
df2 <- read.csv(file.path("data", "3_PeakDischarge.csv"))
# Convert 'Method' column to a factor variable since it's categorical
df2$Method <- as.factor(df2$Method)
```

Boxplot provides an initial insight into whether the variance across groups looks consistent and whether there might be any outliers.

```r
# Boxplot to visualize the distribution of 'PeakDischarge' for each 'Method'
boxplot(PeakDischarge ~ Method, data = df2)
```



In the boxplot of the PeakDischarge data, we observe that the data variation for Methods 3 and 4 appears greater than that for Methods 1 and 2. To confirm whether the variances across the different 'Method' groups are statistically equal, we can perform modified Levene's test for homogeneity of variance. R users can implement the modified Levene's test by installing the package **asbio**. R codes are shown below.

```
# Install the asbio package
# install.packages("asbio")
# Load the asbio package, which includes the modified Levene's test
# for homogeneity of variance
library(asbio)
# Perform Levene's test for homogeneity of variance
# It checks if the variances across the different 'Method' groups are equal
modlevene.test(df2$PeakDischarge, df2$Method)
```

```
##
## Modified Levene's test of homogeneity of variances
##
## df1 = 3,  df2 = 20,  F = 4.56844,  p-value = 0.01357
```

The p-value of the modified Levene's test is 0.0136, which is lower than the pre-specified significance level of 0.05, indicating that the variances across the different 'Method' groups are significantly different.

Now, suppose we did not perform the modified Levene's test and directly fit an one-way ANOVA instead.

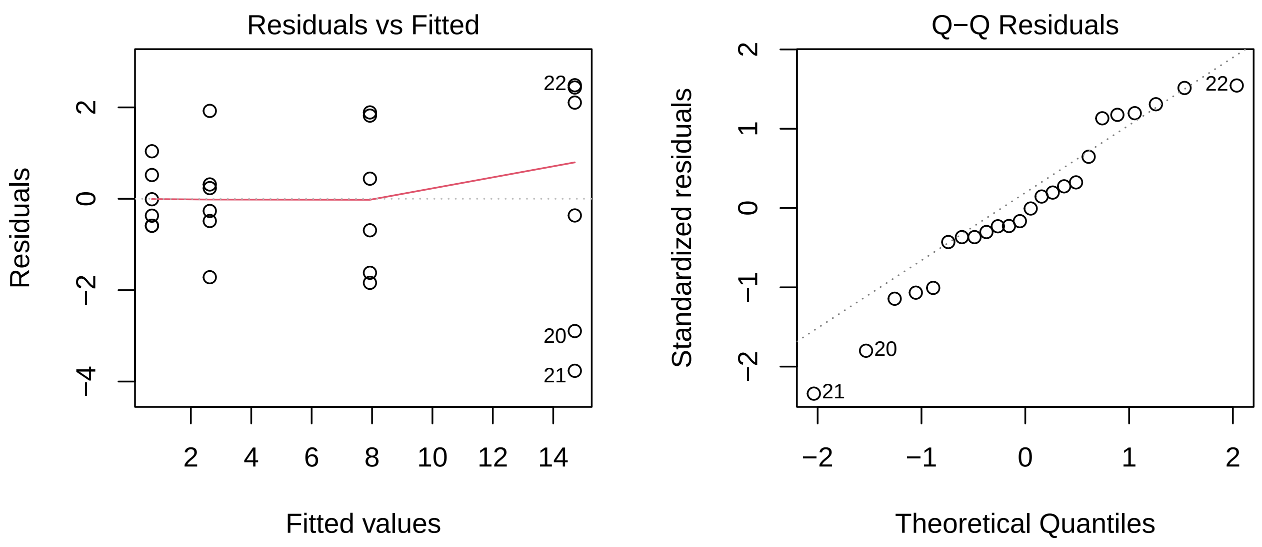
```
# Perform a one-way ANOVA to test if there is a significant difference
# in 'PeakDischarge' across methods
fit2 <- aov(PeakDischarge ~ Method, data = df2)
# Summary of the ANOVA results, including F-statistic and p-value
summary(fit2)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Method       3  708.3   236.1   76.07 4.11e-11 ***
## Residuals   20   62.1     3.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

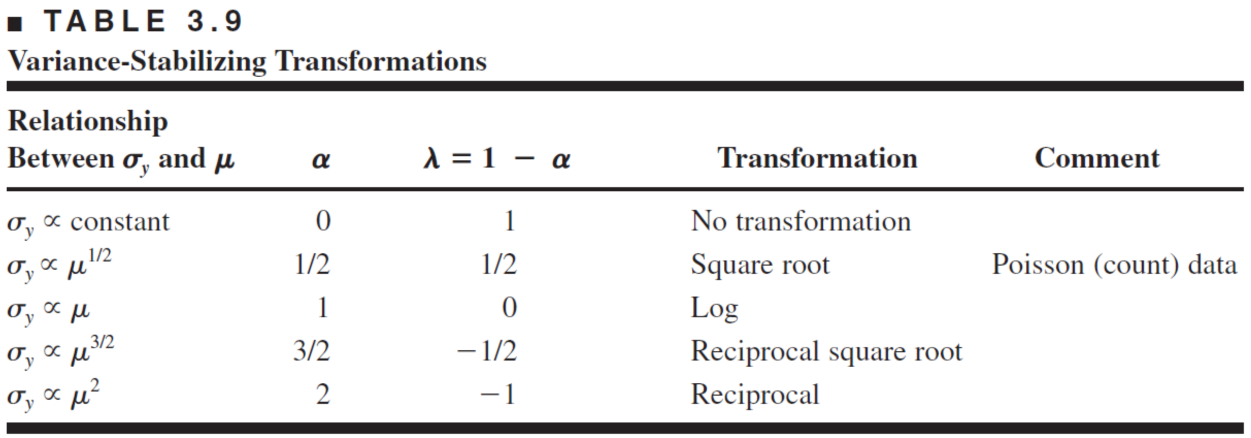
The p-value of the F-test in the ANOVA table is extremely small, providing significant evidence that the mean Peak Discharge differs across the Methods. **But is this conclusion valid?**

If we examine the residual plots, particularly the scatter of fitted values versus residuals, we notice an increasing trend in the variation of residuals across the Method groups. This suggests that the assumption of constant variance in the ANOVA model may be violated.

```r
# Diagnostic plots for the ANOVA model: Residuals and Q-Q plot
par(mfrow = c(1, 2))
plot(fit2, which = 1:2)
```



Referring back to Section 3.4.3, to address the issue of unequal variances among groups, a common approach is to apply the Box-Cox transformation. This technique helps stabilize variances and make the data more normally distributed, aligning with the assumptions of ANOVA. The transformation identifies an optimal parameter, λ (lambda), which suggests the most suitable transformation. The value of λ will guide the appropriate transformation to apply, as outlined in Table 3.9 of the textbook, shown below.



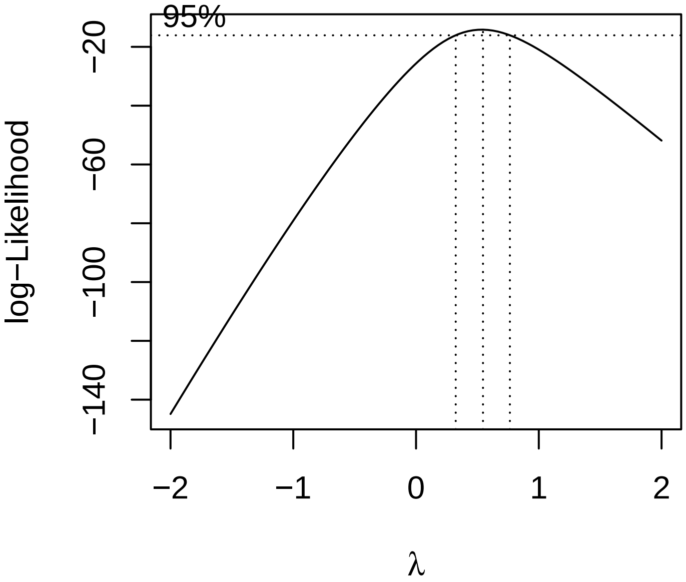
■ **TABLE 3.9**

Variance-Stabilizing Transformations

Relationship Between σ_y and μ	α	$\lambda = 1 - \alpha$	Transformation	Comment
$\sigma_y \propto$ constant	0	1	No transformation	
$\sigma_y \propto \mu^{1/2}$	1/2	1/2	Square root	Poisson (count) data
$\sigma_y \propto \mu$	1	0	Log	
$\sigma_y \propto \mu^{3/2}$	3/2	−1/2	Reciprocal square root	
$\sigma_y \propto \mu^2$	2	−1	Reciprocal	

For R users, there are various ways to implement the Box-Cox transformation. Here, we demostrate the use of the `boxcox` function in the **MASS** package.

```r
# Load the MASS package to apply the Box-Cox transformation
library(MASS)
# Apply the Box-Cox transformation to the ANOVA model to stabilize the variance
fit2bc <- boxcox(fit2, plotit = TRUE)
```



The Box-Cox plot is to visually identify the best transformation (optimal λ) to apply to the data, which will help meet the assumptions of normality and homogeneity of variance in ANOVA model. The x-axis shows different possible values of λ, typically ranging from -2 to 2, depending on the data. The y-axis represents the log-likelihood of the data under different transformations. The goal is to find the value of λ that maximizes the log-likelihood, indicating the optimal transformation to stabilize variance and improve normality.

```r
# Extract the lambda value corresponding to the maximum likelihood estimate
lambda <- fit2bc$x[which.max(fit2bc$y)]
# Print the estimated lambda for the Box-Cox transformation
sprintf("lambda = %.3f", lambda)
```

```
## [1] "lambda = 0.545"
```

From the results generated by the `boxcox` function, we can use the following code to identify the optimal λ value. In this case, the λ value is 0.545, which is close to 0.5. According to Table 3.9 in the textbook, this suggests applying a square root transformation to the response variable `PeakDischarge`.

Before refitting the ANOVA model, we can first perform the modified Levene's test again to check for equal variances of the square root-transformed `PeakDischarge` across the Method groups.

```r
# library(asbio)
# Perform modified Levene's test again after applying the
# square root transformation (common for stabilizing variance)
modlevene.test(sqrt(df2$PeakDischarge), df2$Method)
```

```
##
## Modified Levene's test of homogeneity of variances
##
## df1 = 3,  df2 = 20,  F = 0.23917,  p-value = 0.86798
```

The p-value of the modified Levene's test is 0.868, which is large, indicating that the variances across the different 'Method' groups are likely similar.

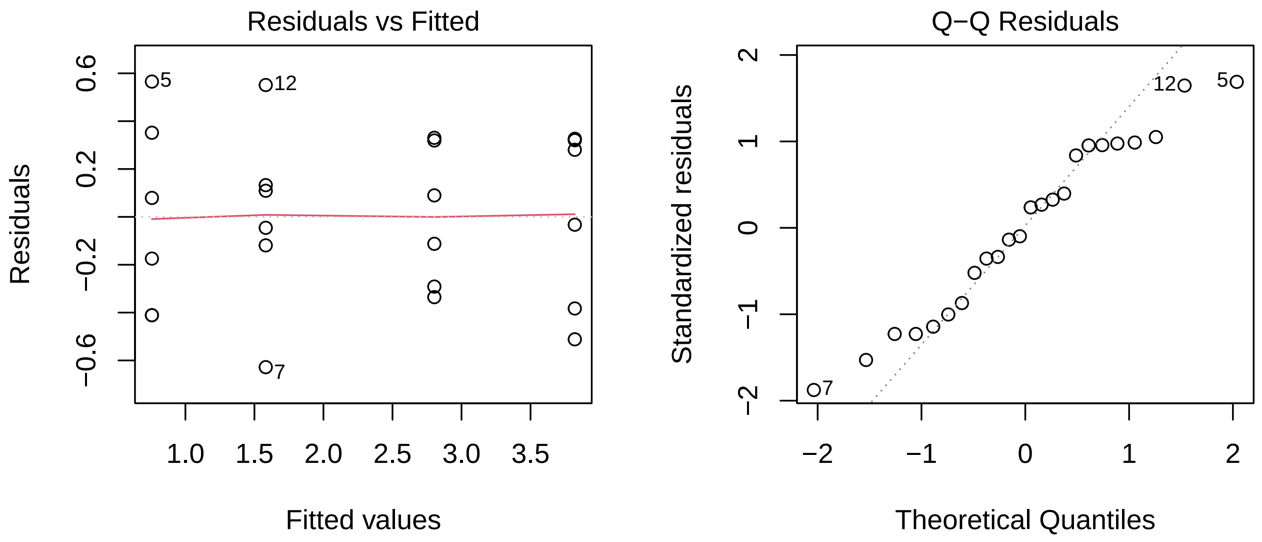
Finally, we fit the ANOVA model using the square root-transformed `PeakDischarge`.

```r
# Fit a new ANOVA model using the square root-transformed 'PeakDischarge'
fit2s <- aov(sqrt(PeakDischarge) ~ Method, data = df2)
# Summary of the new ANOVA results after transformation
summary(fit2s)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)    
## Method        3  32.68  10.895   81.05 2.3e-11 ***
## Residuals    20   2.69   0.134                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the F-test in this new ANOVA table remains extremely small, providing strong evidence that the mean Peak Discharge differs across the Methods. The residual plots confirm that the assumptions of normality and constant variance are satisfied.

```
# Diagnostic plots for the transformed ANOVA model
par(mfrow = c(1, 2))
plot(fit2s, which = 1:2)
```



If there is concern about the normality assumption, as indicated by a light-tailed distribution in the QQ-plot, we can conduct a formal hypothesis test for normality. The goal is to check if the test shows an insignificant result, indicating that the residuals do not deviate significantly from a normal distribution. For example, using the Shapiro-Wilk test for normality, we find that the p-value is 0.414, which is greater than the pre-specified significance level of 0.05. This result indicates there is no evidence to suggest that the residuals are from a non-normal distribution.

```
shapiro.test(fit2s$residuals)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  fit2s$residuals
## W = 0.95877, p-value = 0.4141
```