STA461: ANOVA Diagnostics

Lynn Lin

October 19, 2018

Introduction

- Diagnostics and residuals were discussed in regression
  - The purpose was to make sure the model was appropriate for the data and the data is appropriate for the model
- We will use the same type of approach for ANOVA models
- Just like in regression, the model is pretty robust to minor deviations from the assumptions
  - Just need to look for any major problems

Residuals

- Residual analysis for ANOVA models corresponds closely to that for regression models
- The formula for residual for the ANOVA is

$$e_{it} = Y_{it} - \hat{Y}_{it} = Y_{it} - \bar{Y}_{i.}$$

- Under the assumptions of the one-way ANOVA model, it is normally distributed with mean 0 and variance

$$Var(e_{it}) = \sigma^2(1 - \frac{1}{r_i})$$

- e_{it}'s are sometimes called *raw* residuals to distinguish from the standardized residuals

Standardized residuals

- Replacing σ with $\sqrt{MSE}$, we define the **standardized residual** as

$$sr_{it} = \frac{Y_{it} - \bar{Y}_{i.}}{\sqrt{MSE}\sqrt{1 - \frac{1}{r_i}}}$$

- The variance of sr_{it} is approximately one
  - It is not exactly one, because σ has been replaced by $\sqrt{MSE}$
  - Its distribution is not exactly Student's t, because the numerator and denominator aren't independent
- When $n - \nu$ is large, the standardized residuals can be treated as approximately independent and standard normal

Diagnosis of departures from ANOVA model

Residual plots can be helpful in diagnosing the following departures from ANOVA model:

1. Nonconstancy of error variance
   - The variance may change as a function of the factor level
2. Nonindependence of error terms
   - Some relationship among the error terms
3. Outliers
4. Omission of important explanatory variables
   - One or several important factors have been omitted from the model
5. Nonnormality of error terms

Nonconstancy of error variance

ANOVA model requires the error terms ϵ_{it} have constant variance for all factor levels

- Test this by plotting of standardized residual against fitted value (corrects for non-equal sample sizes across groups)
- Or boxplots of the standardized residuals for each treatment
- Under the constancy of the error variance assumption, the plot should have the same extent of scatter of the residuals around zero for each factor level

Example

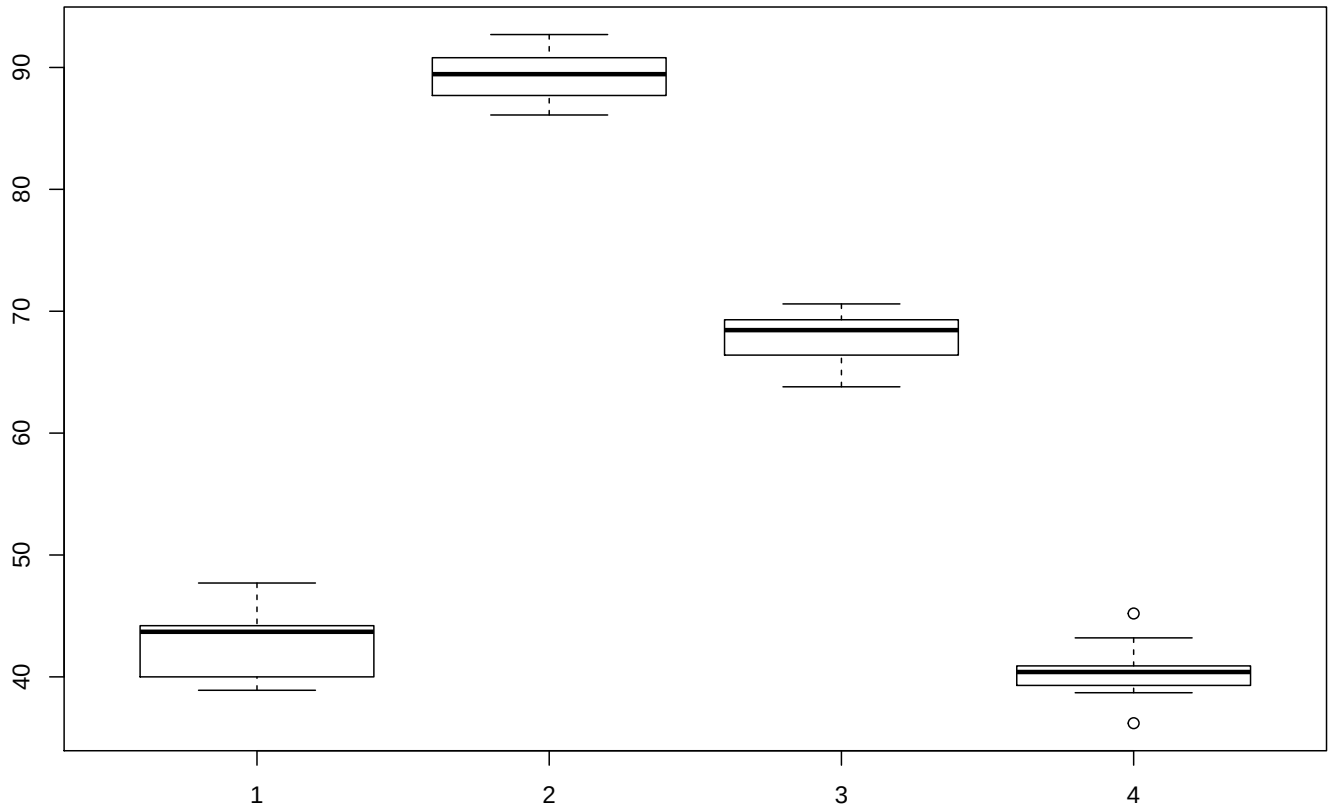
- To study the effectiveness of different rust inhibitors, four brands (1,2,3,4) were tested
- 40 experimental units were randomly assigned to the four brands with 10 units assigned to each brand

```
rust$brand <- as.factor(rust$brand)
rust.aov <- aov(Y ~ brand, data = rust)
anova(rust.aov)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## brand      3  15954  5317.8  866.12 < 2.2e-16 ***
## Residuals 36    221     6.1
## ---
```

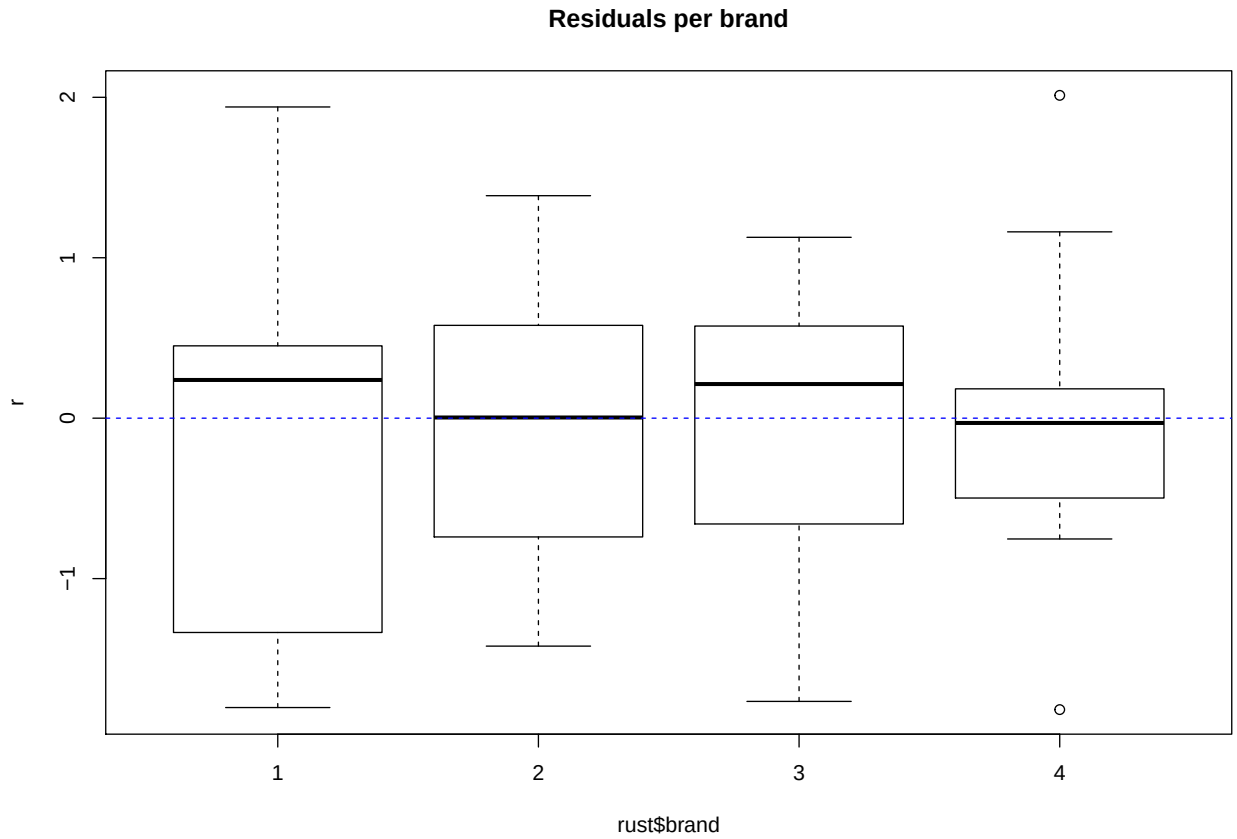
Example

```
boxplot(rust$Y ~ rust$brand)
```



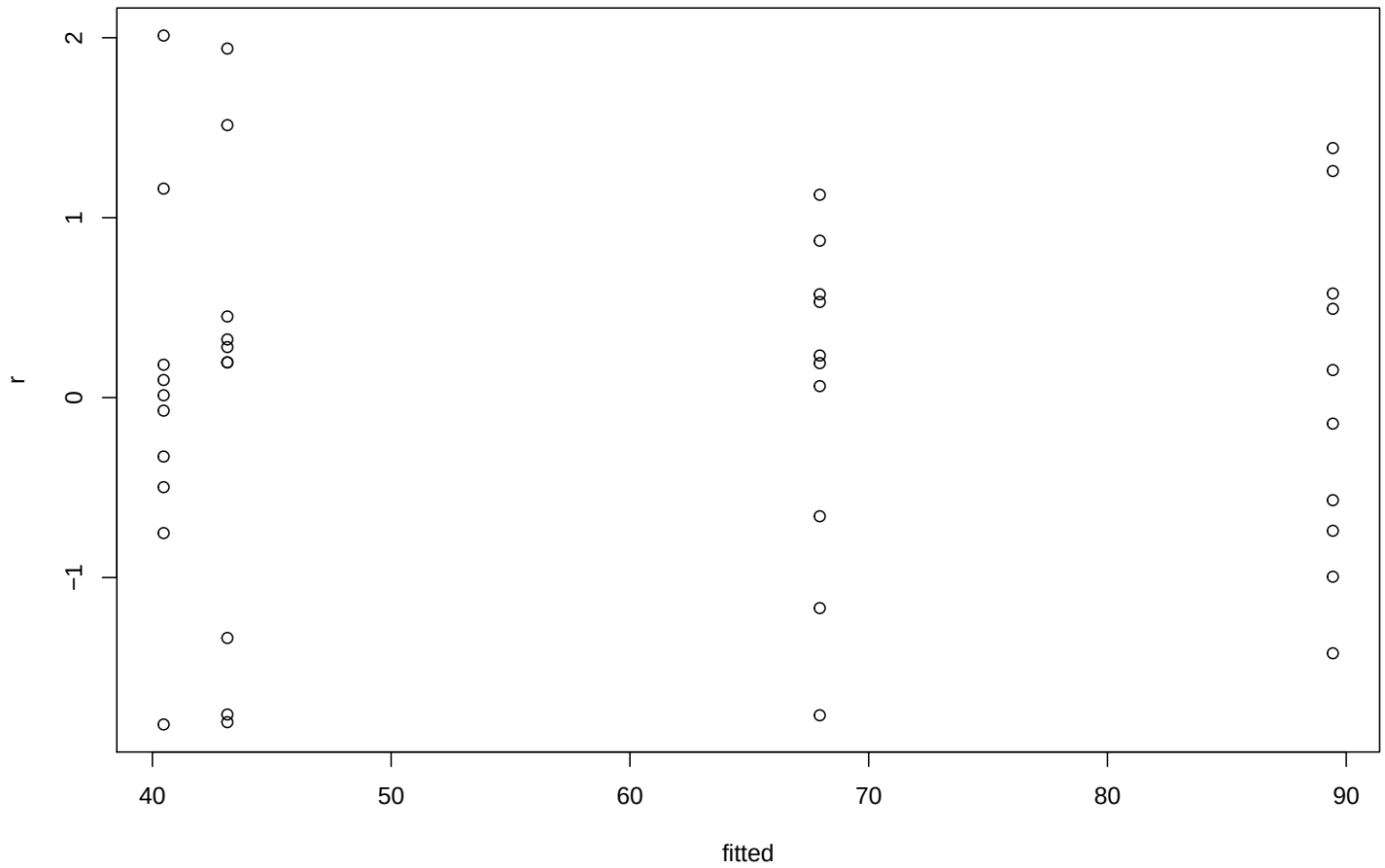
Example

```
r <- rstandard(rust.aov)
plot(r ~ rust$brand, main = "Residuals per brand")
abline(0, 0, lty = 2, col = "blue")
```



Nonconstancy of error variance

```
fitted = fitted(rust.aov)
plot(fitted, r)
```



Nonindependence of error terms

- Appropriate to test when there is some time sequence involves
- If there is no time factor or some ordering sequence, you should pretty much assume independence, unless you have some underlying assumption of the data
    - E.g., the data are ordered in some logical sequence, such as in a geographic sequence
- Test this by ploting residuals across time, plotting residuals across time by each factor

Outliers

- Outliers are single observations whose residuals are unusually large
- To detect outliers:
  - Residuals against fitted values
  - Residual dot plots
- If an observation is truly unusual, and there is a reason to think that this unit should behave differently from the others in the same group, then removing it may be appropriate
- Hypothesis-testing procedures for outliers are available, but these tests usually assume that the population is homoscedastic and normal, and departures from that assumption will dramatically affect the performance of the tests

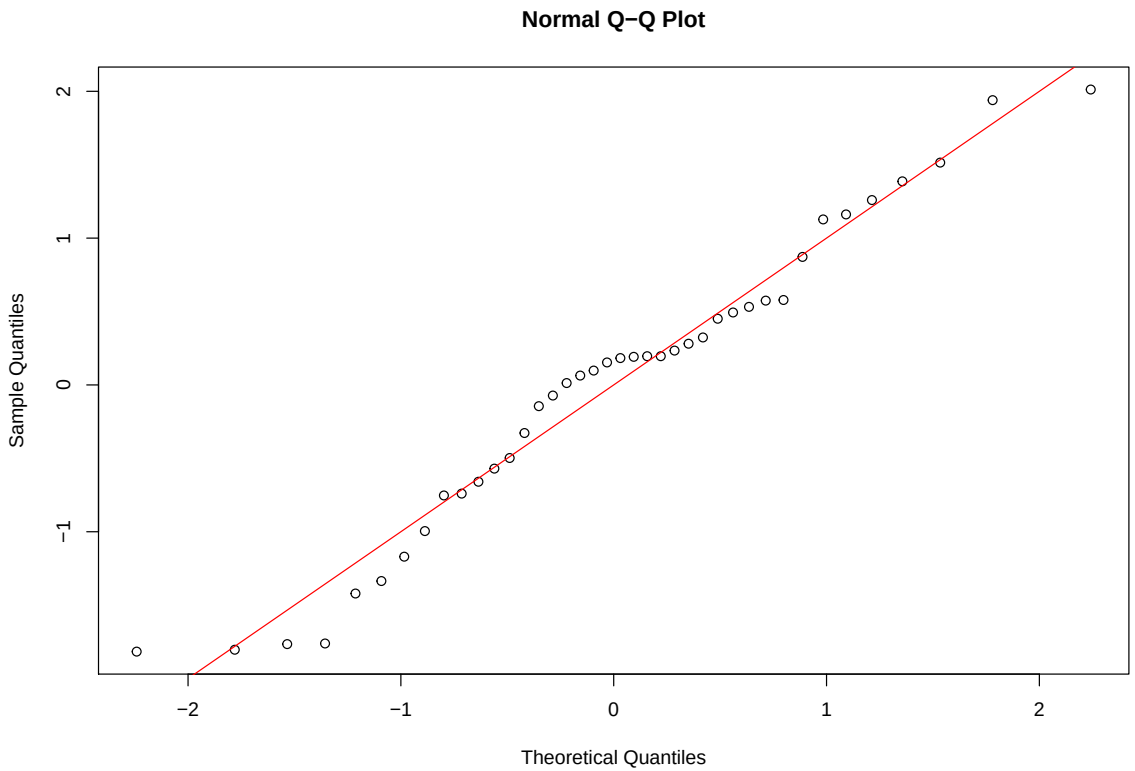
Omission of an important factor

- You want to make sure you didn't leave out an important factor that helps describe the data
- Plot residuals by fitted values by the left out factor
  - Test to see if the residuals are grouped by the factor levels

Nonnormality of error terms

- Use normal probably plot of the residuals

```
qqnorm(r)
abline(0, 1, col = "red")
```



Transformations of response variable

When the model assumptions of constancy of the error variance and/or normality of the error distributinos are violated, a transformation of the response variable is often useful

- Some simple guides
- Box-Cox procedure

Some simple guides

Four simple guides to finding a useful transformation:

- When σ_i^2 is proportional to μ_i, a square root transformation is helpful
  - $Y' = \sqrt{Y}$
  - $Y' = \sqrt{Y} + \sqrt{Y+1}$
  - This often happens when Y is a count
- When σ_i is proportional to μ_i, the logarithmic transformation is helpful
  - $Y' = \log(Y)$
- When σ_i is proportional to μ_i^2, the reciprocal transformation is helpful
  - $Y' = \frac{1}{Y}$
- When Y_{it} is a proportion p_{it}, the arcsine transformation is helpful
  - $Y' = 2arcsin\sqrt{Y}$

Use of simple guides

To examine whether one of the simple transformation guides is applicable, some statistics should be calculated for each factor level:

- $s_i^2 / \bar{Y}_{i.}$
- $s_i / \bar{Y}_{i.}$
- $s_i^2 / \bar{Y}_{i.}^2$

Approximate constancy of one of the above three statistics over all factor levels would suggest the corresponding transformation as useful for stablizing the error variance and making the error distributions more nearly normal

Box-Cox procedure

- Box-Cox transformation tries to identify a power transformation of the type of Y^{λ} to correct for both lack of normality and nonconstant of the error variance
- It numerically search for the optimal λ such that the resulting fitted model has the smallest SSE

Example

```
data <- data.frame(Y = c(0.178, 0.195, 0.225, 0.294, 0.315, 0.
    0.398, 0.407, 0.409, 0.432, 0.494, 0.719, 0.11, 0.111, 0.2
    0.492, 0.965, 1.113, 1.19, 1.233, 1.505, 1.897, 0.106, 0.1
    0.51, 0.576, 0.588, 0.608, 0.64, 0.658, 0.788, 0.958), GP
    13), rep("C", 13)))
head(data)
```

```
##       Y GP
## 1 0.178  A
## 2 0.195  A
## 3 0.225  A
## 4 0.294  A
## 5 0.315  A
## 6 0.341  A
```

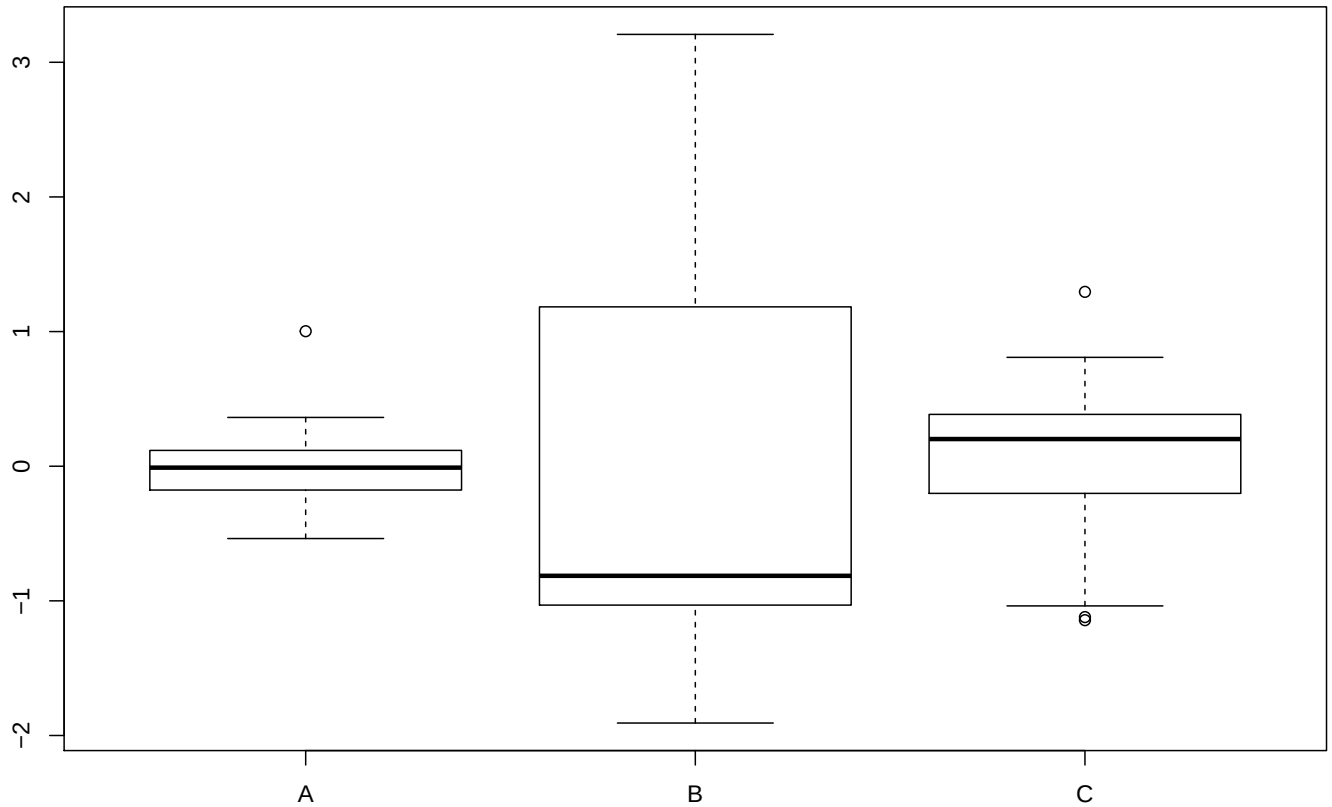
Example

```
with(data, tapply(Y, GP, var))
```

```
##           A           B           C
## 0.01734578 0.33182844 0.06673060
```

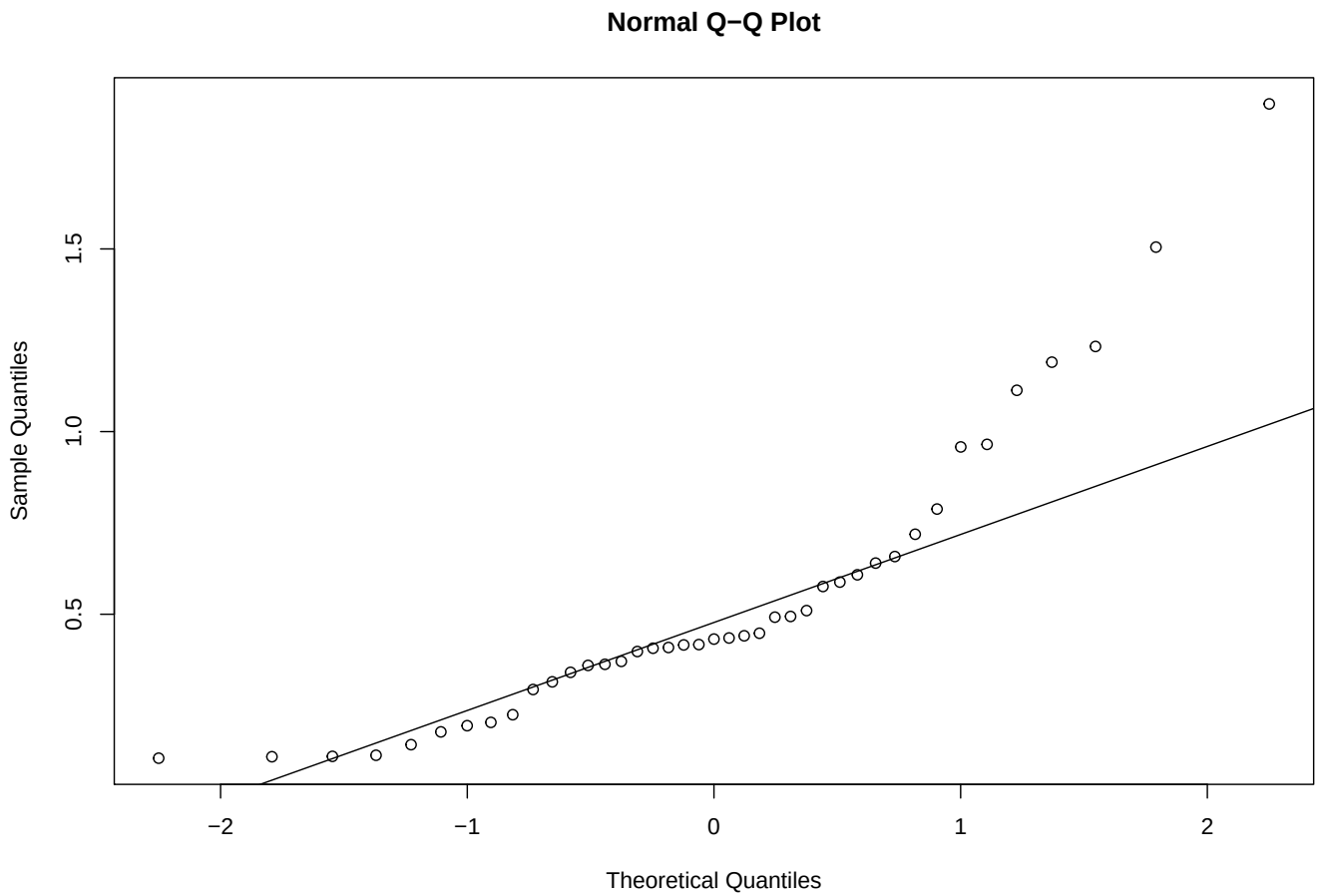
Example

```
m1 <- aov(Y ~ GP, data)
plot(data$GP, rstandard(m1))
```



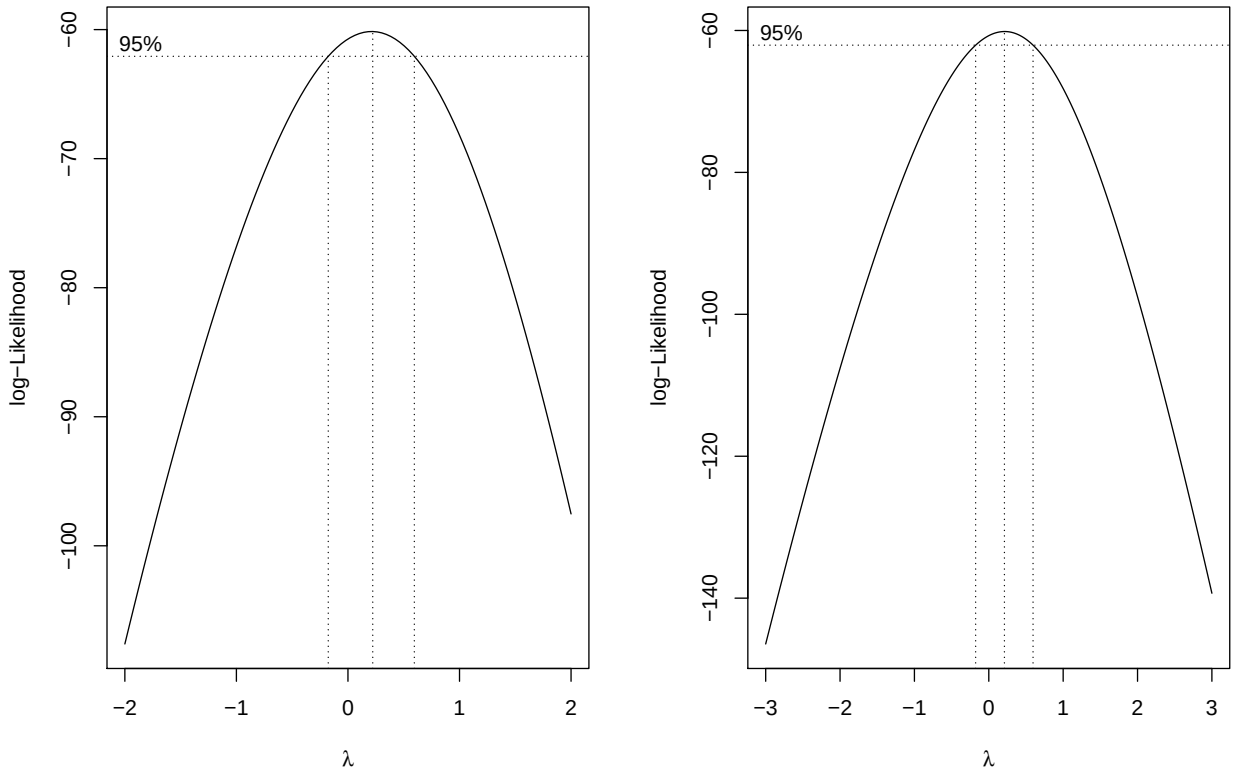
Example

```
qqnorm(data$Y)
qqline(data$Y)
```



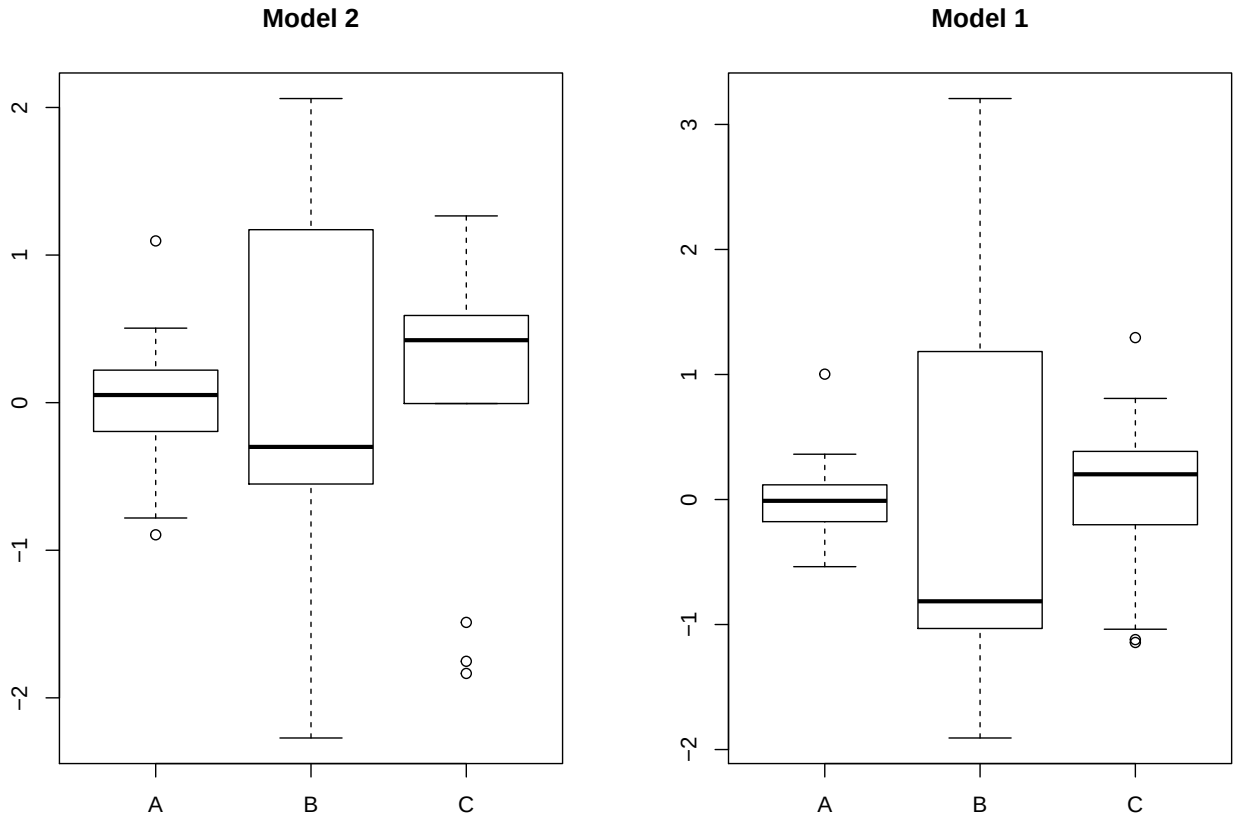
Example

```
library(MASS)
par(mfrow = c(1, 2))
boxcox(m1)
boxcox(m1, lambda = seq(-3, 3, by = 0.01))
```



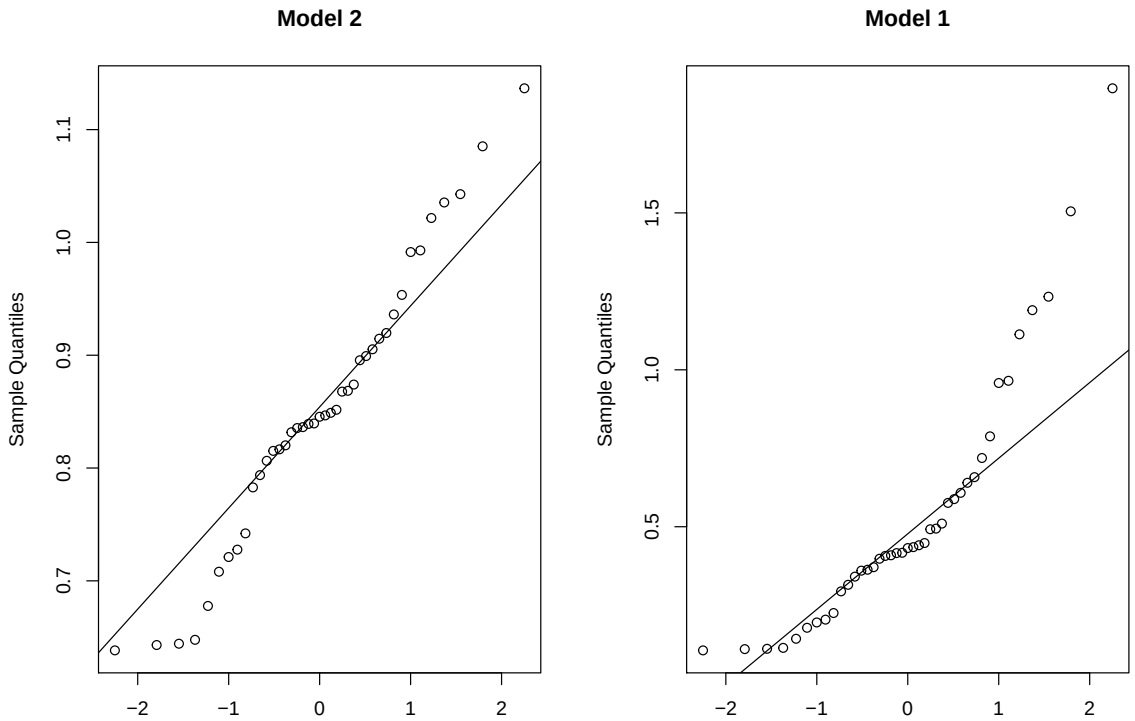
Example

```
m2 <- aov(Y^0.2 ~ GP, data)
par(mfrow = c(1, 2))
plot(data$GP, rstandard(m2), main = "Model 2")
plot(data$GP, rstandard(m1), main = "Model 1")
```



Example

```
par(mfrow = c(1, 2))
qqnorm(data$Y^0.2, main = "Model 2")
qqline(data$Y^0.2)
qqnorm(data$Y, main = "Model 1")
qqline(data$Y)
```



Effects of departures from model

Nonnormality

- The point estimators of factor level means and linear combinations are unbiased whether or not the populations are normal
- F test for equality of factor level means is a robust test against departures from normality

Unequal error variances

- The F test and related analyses are robust against unequal variances when the sample sizes are approximately equal
  - F test for equality of factor level means
  - Scheffe multiple comparison procedure based on the F distribution
- Single comparions between factor level means can be substantially affected by unequal variances

Nonindependence of error terms

- Lack of independence of the error terms can have serious effects on inferences in the analysis of variance
- Modify the model

Example 2

Three brands of batteries are under study. It is suspected that the lives (in weeks) of the three brands are different. Five batteries of each brand are tested with the following results:

```
##       y brand
## 1   100     1
## 2    96     1
## 3    92     1
## 4    96     1
## 5    92     1
## 6    76     2
## 7    80     2
## 8    75     2
## 9    84     2
## 10   82     2
## 11  108     3
```

Example 2

Are the lives of these brands of batteries different?

```
fit <- aov(y ~ brand, data)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## brand        2 1196.1   598.1   38.34 6.14e-06 ***
## Residuals   12  187.2    15.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

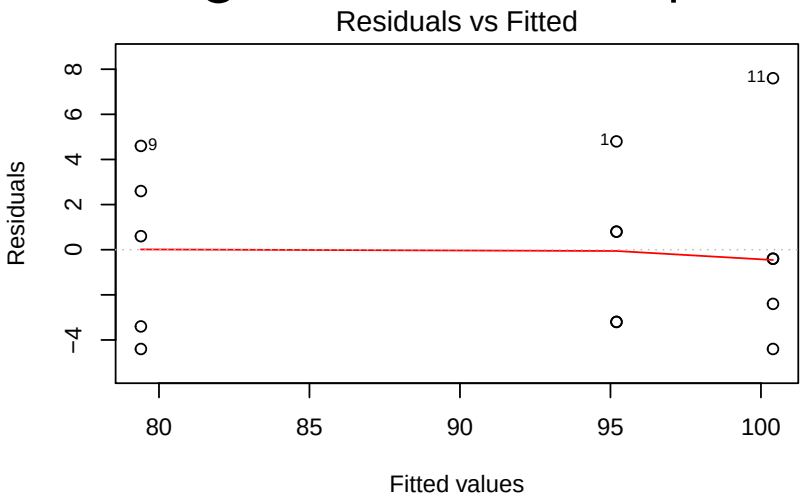
Example 2

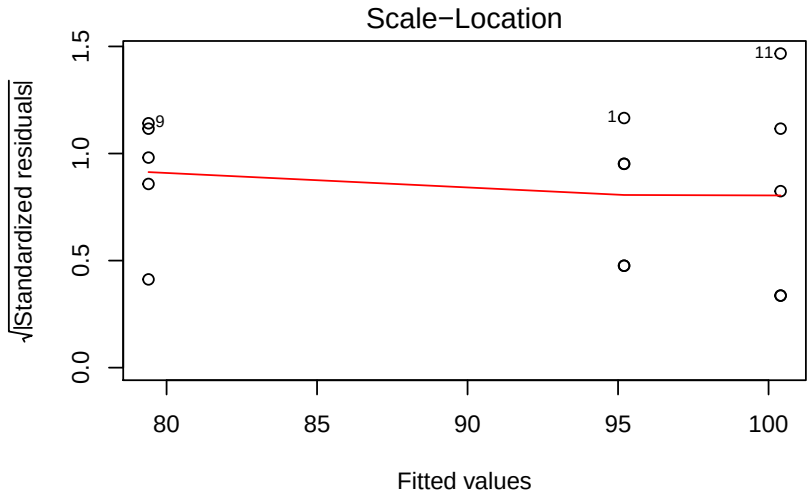
Checking for model assumption

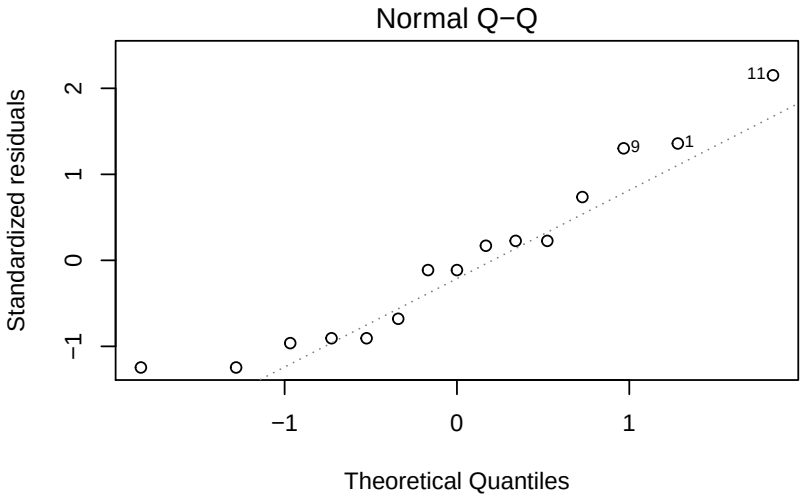
```
layout(matrix(c(1, 2, 3, 4), 2, 2))
plot(fit)
```

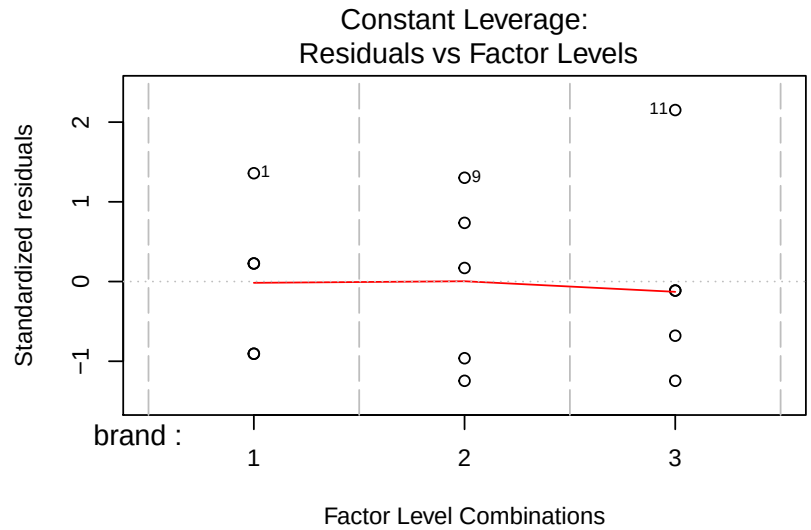
Example 2

Checking for model assumption





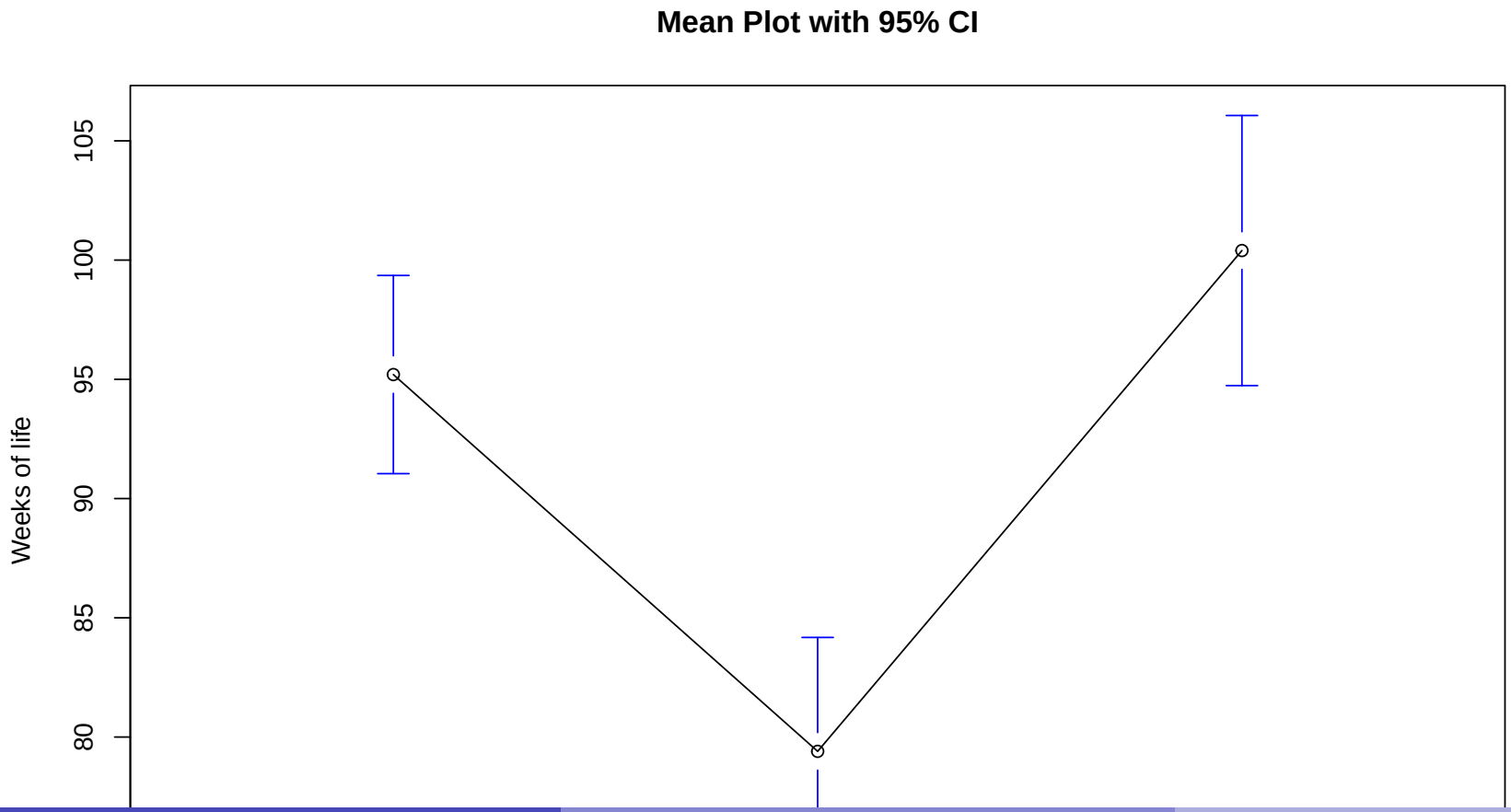




Example 2

Which brand would you select for use?

```
library(gplots)
plotmeans(y ~ brand, xlab = "brands", ylab = "Weeks of life",
```



Example 3

An experiment was performed to investigate the effectiveness of five insulating materials. Four samples of each material were tested at an elevated voltage level to accelerate the time to failure. The failure times (in minutes) is shown below.

```
##         y material
## 1     110        1
## 2     157        1
## 3     194        1
## 4     178        1
## 5       1        2
## 6       2        2
## 7       4        2
## 8      18        2
## 9     880        3
## 10   1256        3
```

Example 3

Do all five materials have the same effect on mean failure time?

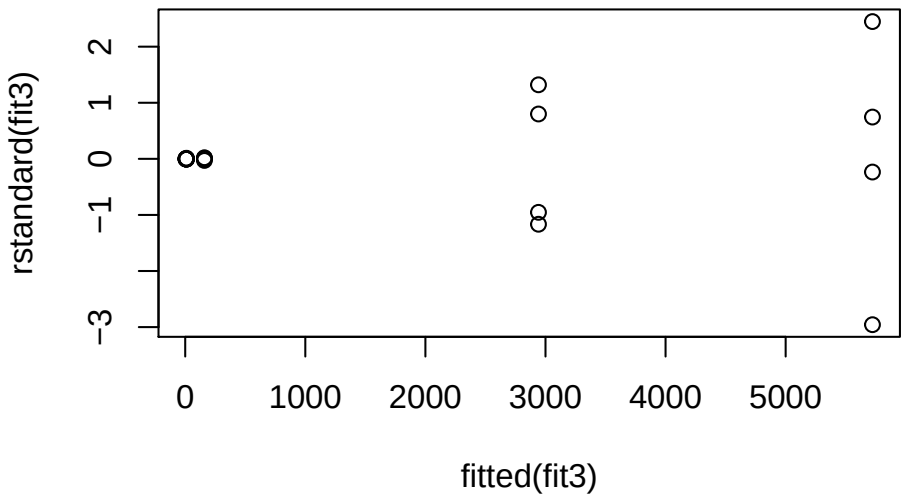
```
fit3 <- aov(y ~ material, data)
summary(fit3)
```

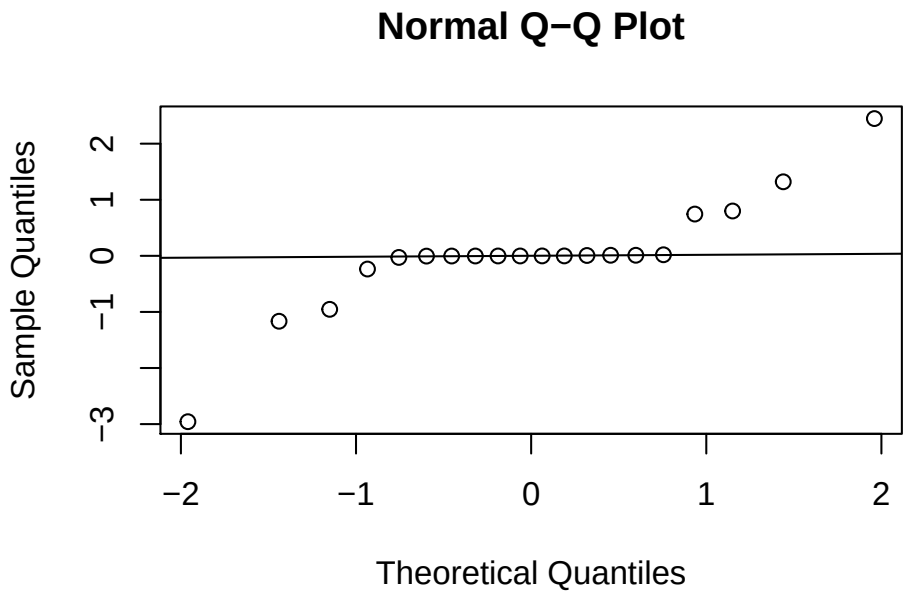
```
##              Df    Sum Sq  Mean Sq F value  Pr(>F)
## material      4 103191489 25797872   6.191 0.00379 **
## Residuals    15  62505657  4167044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Example 3

Plot the residuals versus the predicted response. Construct a normal probability plot of the residuals. What information do these plots convey?

```
par(mfrow = c(1, 2))
plot(fitted(fit3), rstandard(fit3))
qqnorm(rstandard(fit3))
qqline(rstandard(fit3))
```

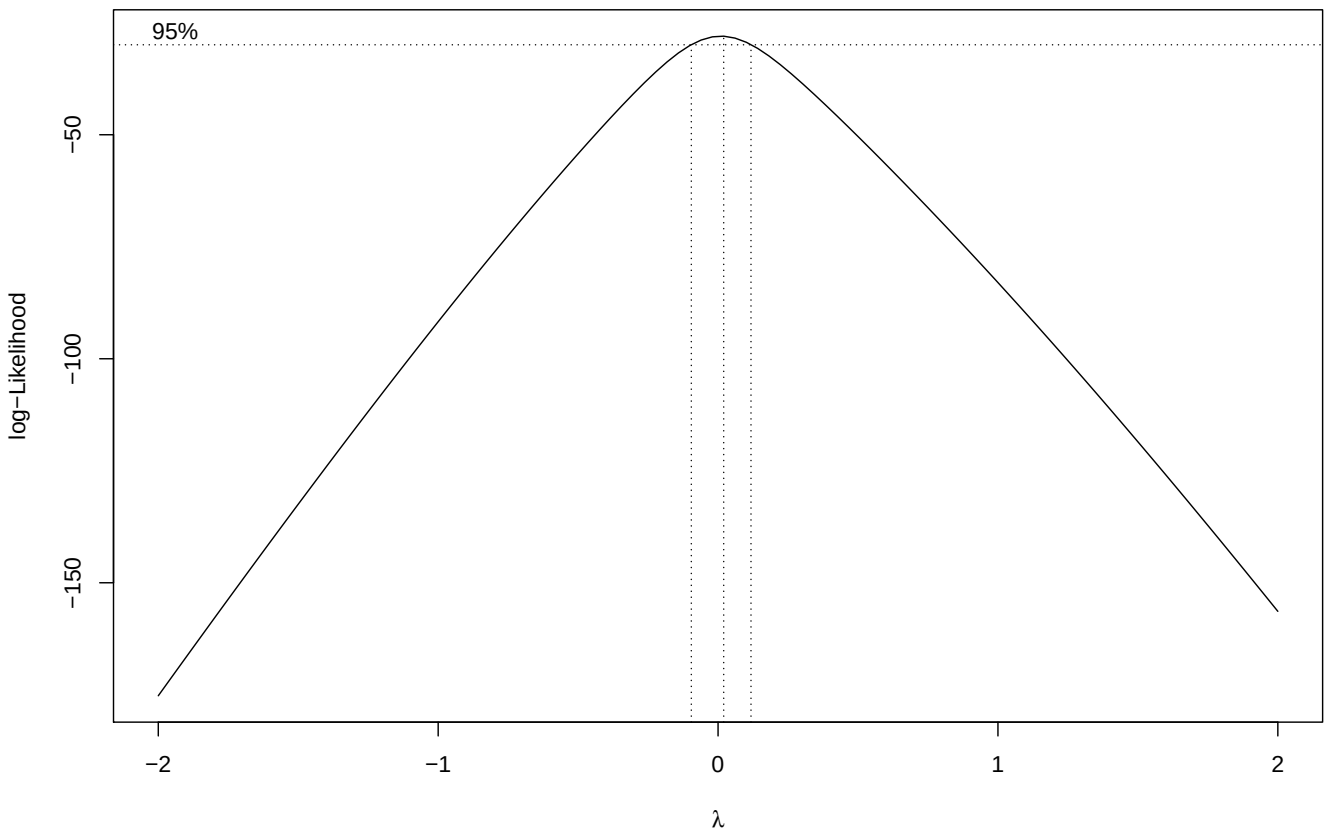




Example 3

Data transformation

```
boxcox(fit3)
```



Example 3

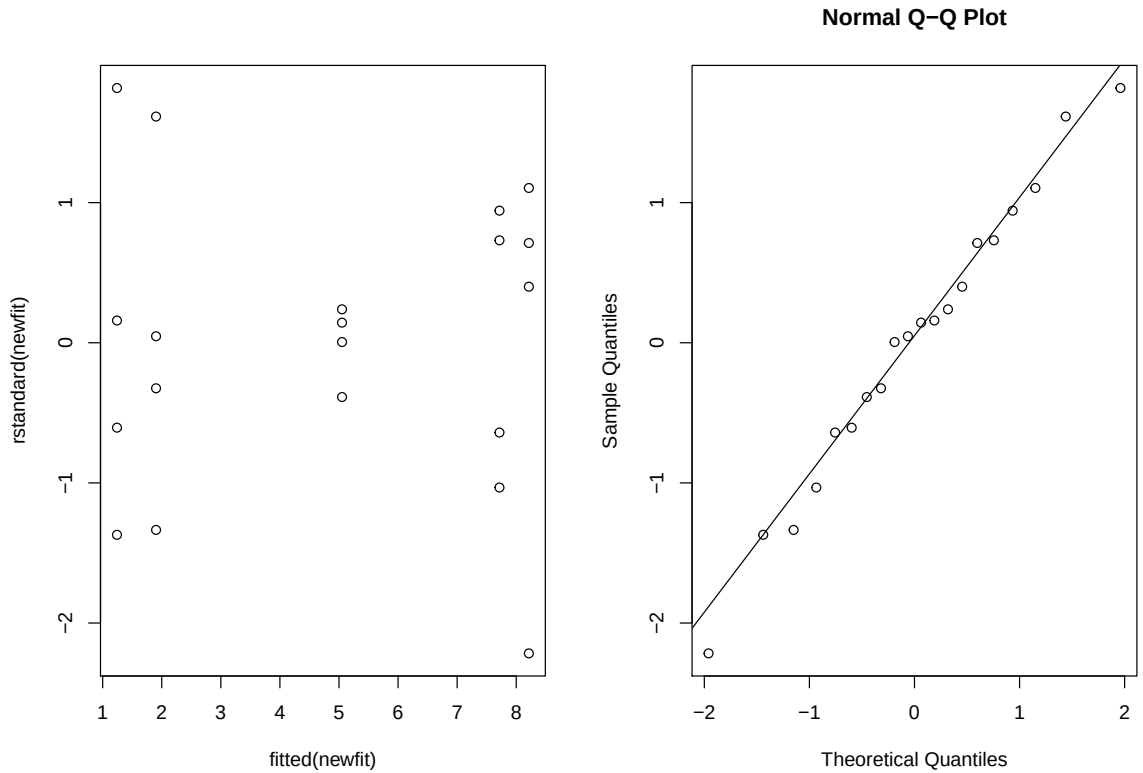
Conduct another analysis of the failure time data

```
newfit <- aov(log(y) ~ material, data)
summary(newfit)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## material      4 165.06   41.26   37.66 1.18e-07 ***
## Residuals    15  16.44    1.10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Example 3

```
par(mfrow = c(1, 2))
plot(fitted(newfit), rstandard(newfit))
qqnorm(rstandard(newfit))
qqline(rstandard(newfit))
```



Example 4

An experiment was performed to investigate the effectiveness of five insulating materials. Four samples of each material were tested at an elevated voltage level to accelerate the time to failure. The failure times (in minutes) is shown below.

```
##       y method
## 1    31      1
## 2    10      1
## 3    21      1
## 4     4      1
## 5     1      1
## 6    62      2
## 7    40      2
## 8    24      2
## 9    30      2
## 10   35      2
```

Example 4

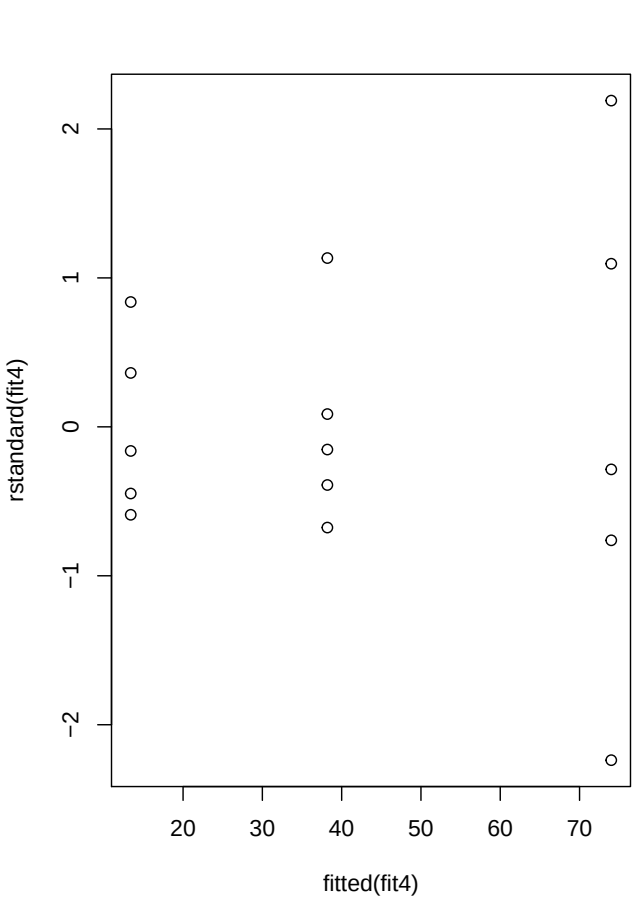
Do all methods have the same effect on mean particle count?

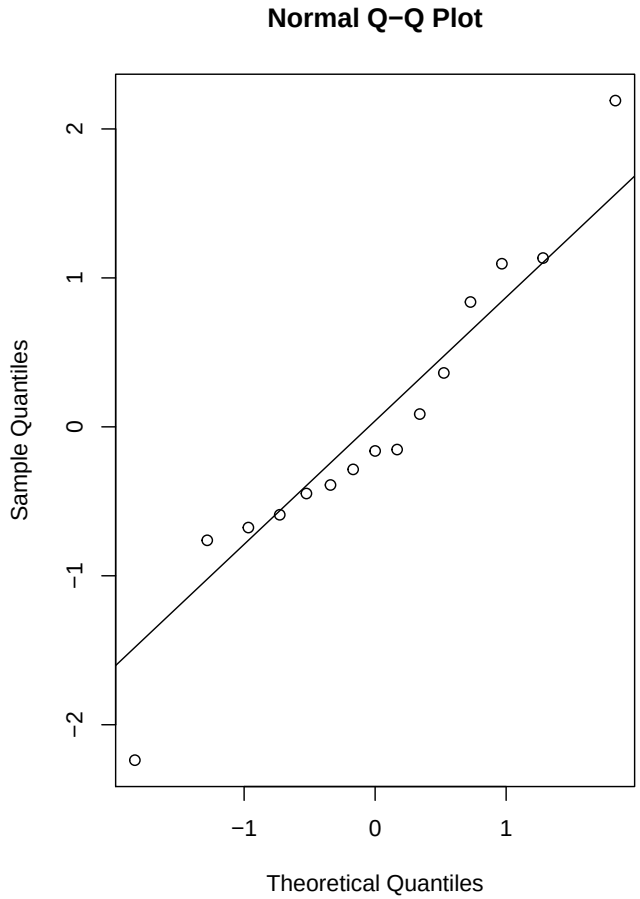
```
fit4 <- aov(y ~ method, data)
summary(fit4)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## method        2   9282    4641   8.418 0.00519 **
## Residuals    12   6616     551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Example 4

Diagnostic plots

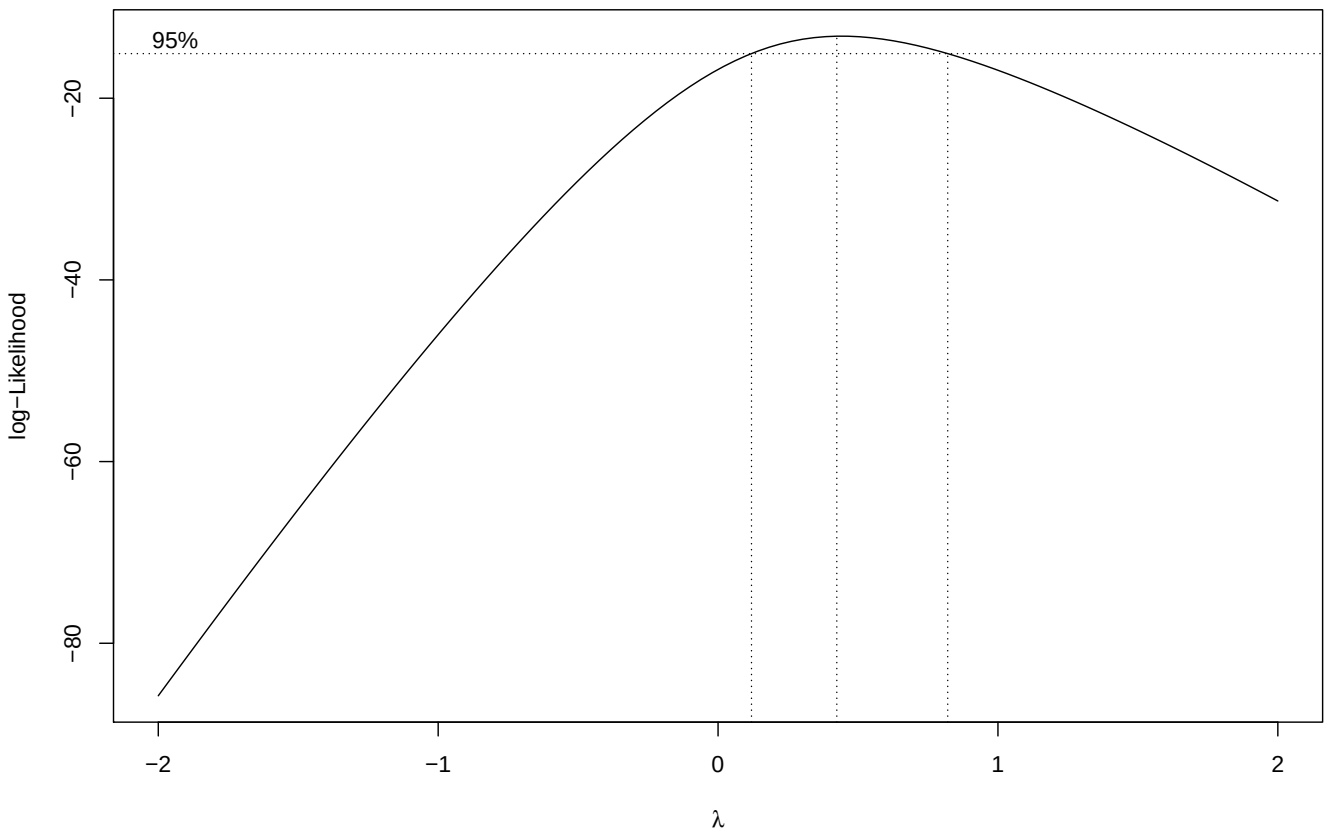




Example 4

Transformation: square-root for count data

```
boxcox(fit4)
```



Example 4

```
newfit <- aov(sqrt(y) ~ method, data)
summary(newfit)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## method        2  65.54   32.77   10.25 0.00253 **
## Residuals    12  38.37    3.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Example 4

```
par(mfrow = c(1, 2))
plot(fitted(newfit), rstandard(newfit))
qqnorm(rstandard(newfit))
qqline(rstandard(newfit))
```