Anova

DATA621 Blog 04

Zhi Ying Chen

27 November 2020

ANOVA (ANalysis Of VAriance) is a statistical test to determine whether two or more population means are different. In other words, it is used to compare two or more groups to see if they are significantly different.

Although ANOVA is used to make inference about means of different groups, the method is called “analysis of variance”. It is called like this because it compares the “between” variance (the variance between the different groups) and the variance “within” (the variance within each group). If the between variance is significantly larger than the within variance, the group means are declared to be different. Otherwise, we cannot conclude one way or the other. The two variances are compared to each other by taking the ratio (variancebetweenvariancewithin) and then by comparing this ratio to a threshold from the Fisher probability distribution (a threshold based on a specific significance level, usually 5%).

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3

## -- Attaching packages -------------------------------------------------------------- tidyverse 1.3.0

## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'readr' was built under R version 3.5.3

## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## Warning: package 'forcats' was built under R version 3.5.3

## -- Conflicts ----------------------------------------------------------------- tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

```
data(InsectSprays)
str(InsectSprays)
```

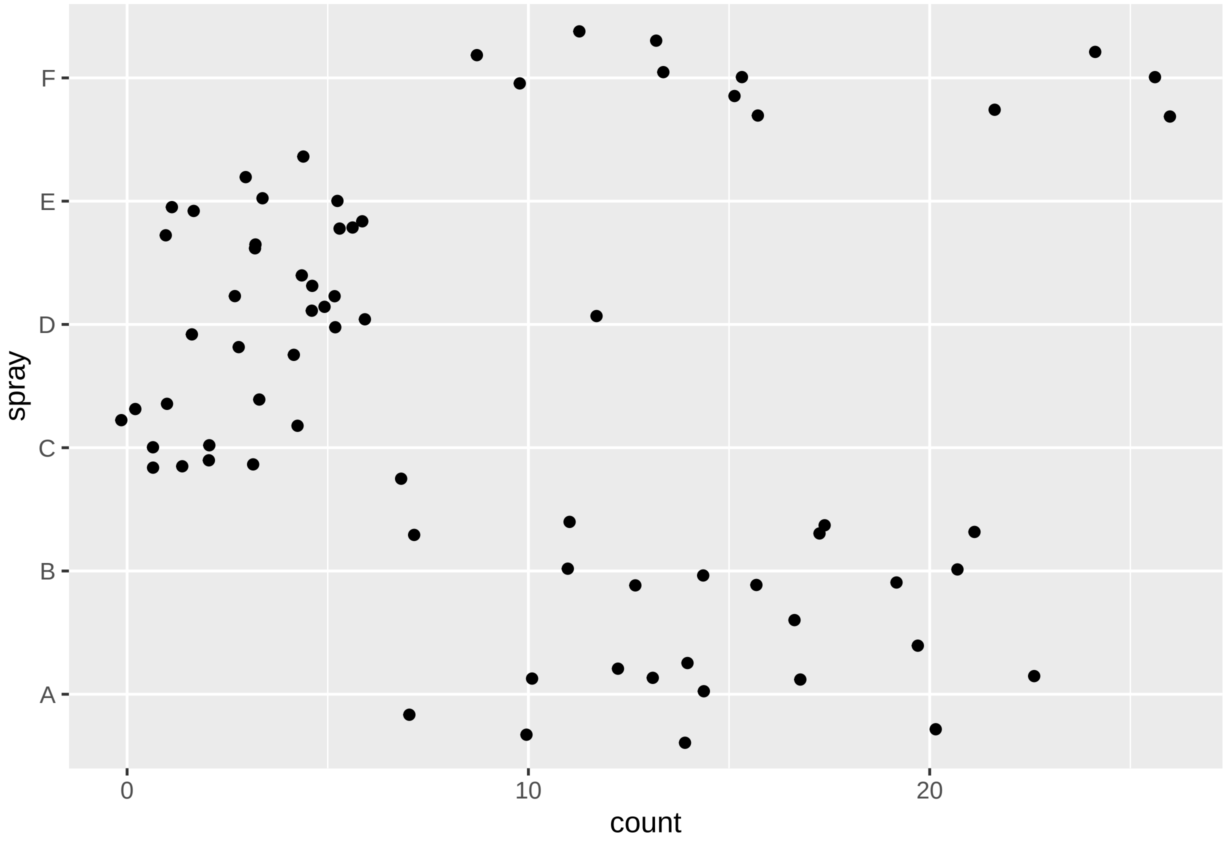
```
## 'data.frame':    72 obs. of  2 variables:
##  $ count: num  10 7 20 14 14 12 10 23 17 20 ...
##  $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(InsectSprays)
```

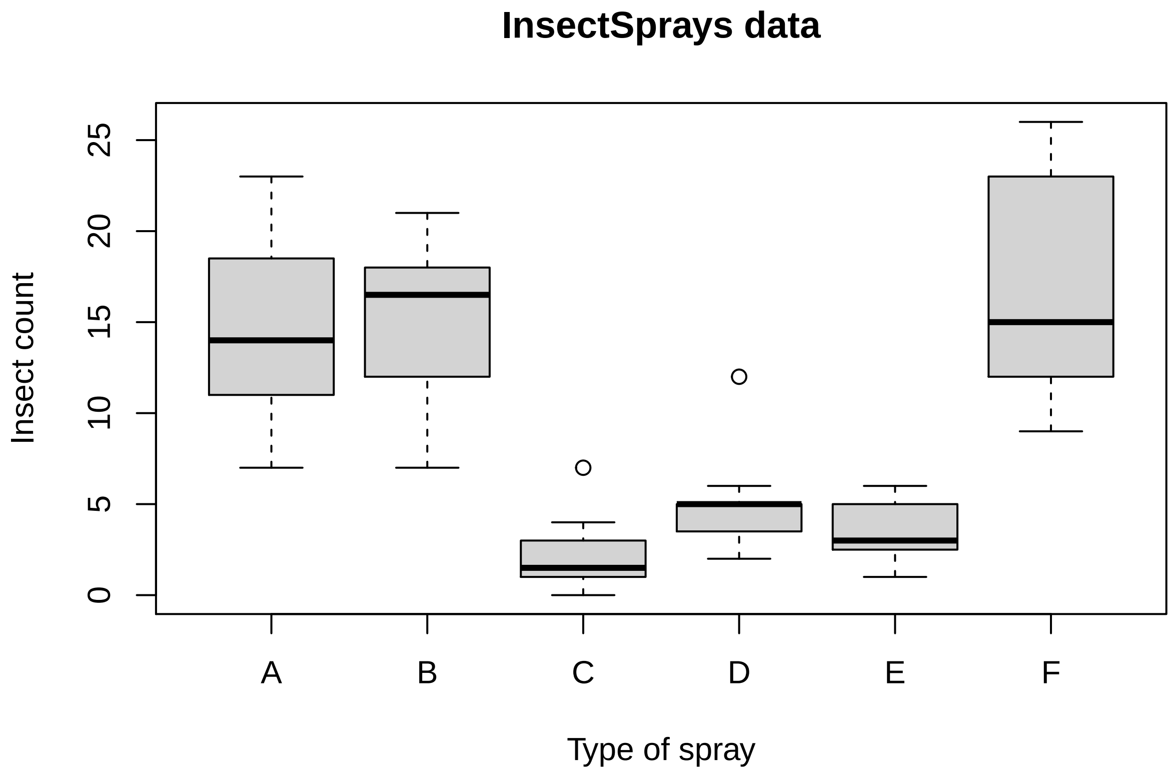
```
##      count        spray
##  Min.   : 0.00   A:12
##  1st Qu.: 3.00   B:12
##  Median : 7.00   C:12
##  Mean   : 9.50   D:12
##  3rd Qu.:14.25   E:12
##  Max.   :26.00   F:12
```

A data frame with 72 observations on 2 variables.

```
ggplot(InsectSprays) +
  aes(x = count, y = spray) +
  geom_jitter() +
  theme(legend.position = "none")
```



```
boxplot(count ~ spray, data = InsectSprays,
        xlab = "Type of spray", ylab = "Insect count",
        main = "InsectSprays data", varwidth = TRUE, col = "lightgray")
```



First Model: Run One-way ANOVA in R

```
oneway.test(count~spray, data=InsectSprays)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  count and spray
## F = 36.065, num df = 5.000, denom df = 30.043, p-value = 7.999e-12
```

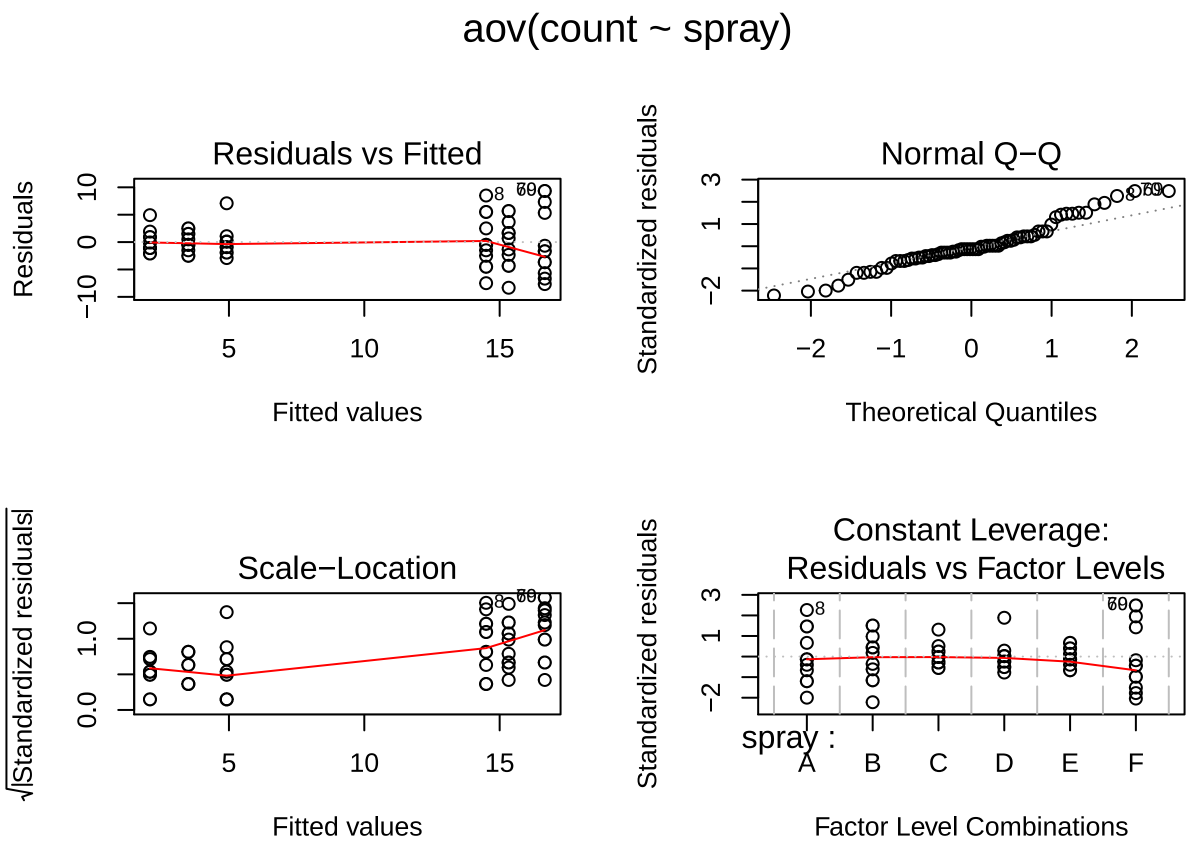
Default is equal variances not assumed that is Welch's correction applied and this explains why the denom df (which is k*{n-1}) is not a whole number in the output O. Oneway.test( ) corrects the non-homogeneity but doesn't give much information. So we only got F score as 36.065 and p-value is 7.999e-12.

Second Model: Run an ANOVA using aov( )

```
Anova_Output <- aov(count ~ spray, data=InsectSprays)
summary(Anova_Output)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## spray        5   2669   533.8    34.7 <2e-16 ***
## Residuals   66   1015    15.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
opar <- par(mfrow = c(2, 2), oma = c(0, 0, 1.1, 0))
plot(Anova_Output)
```

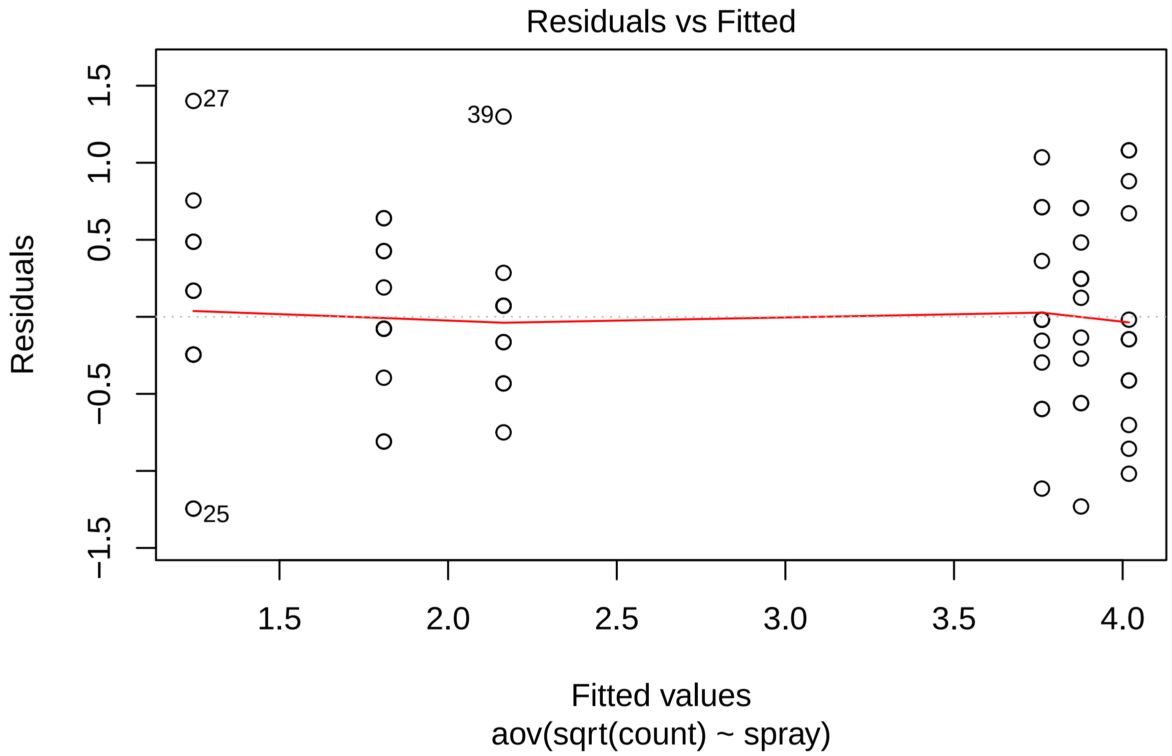


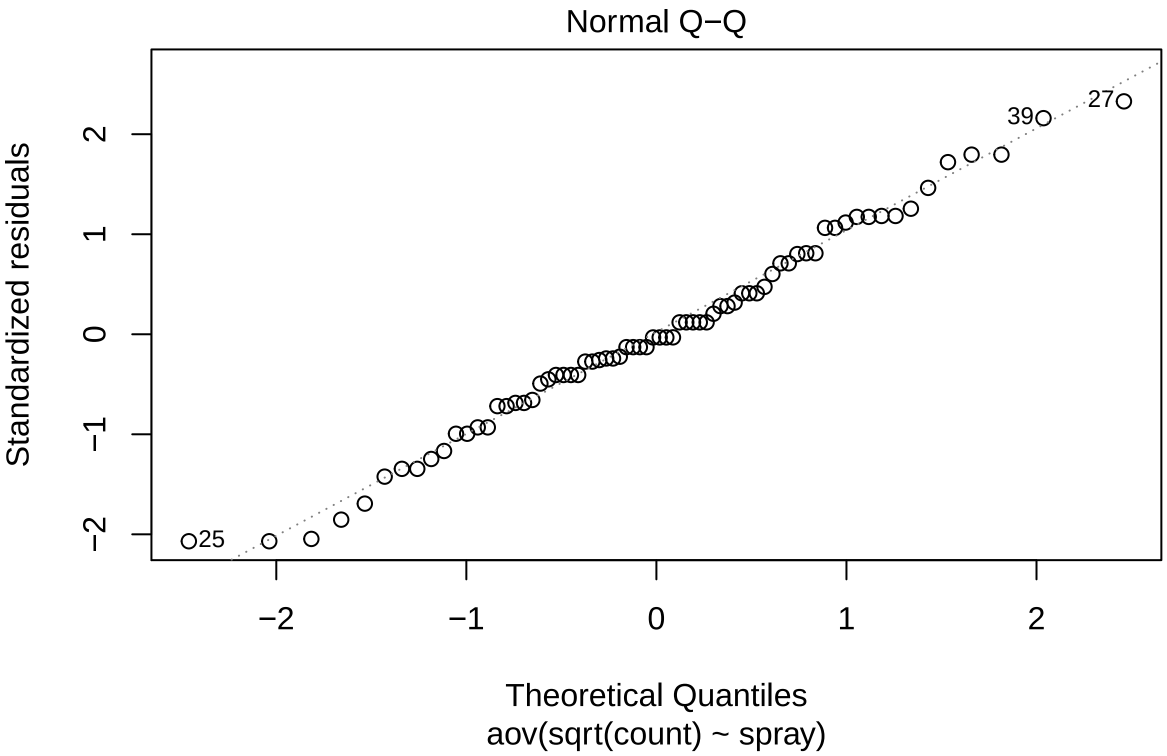
Third model

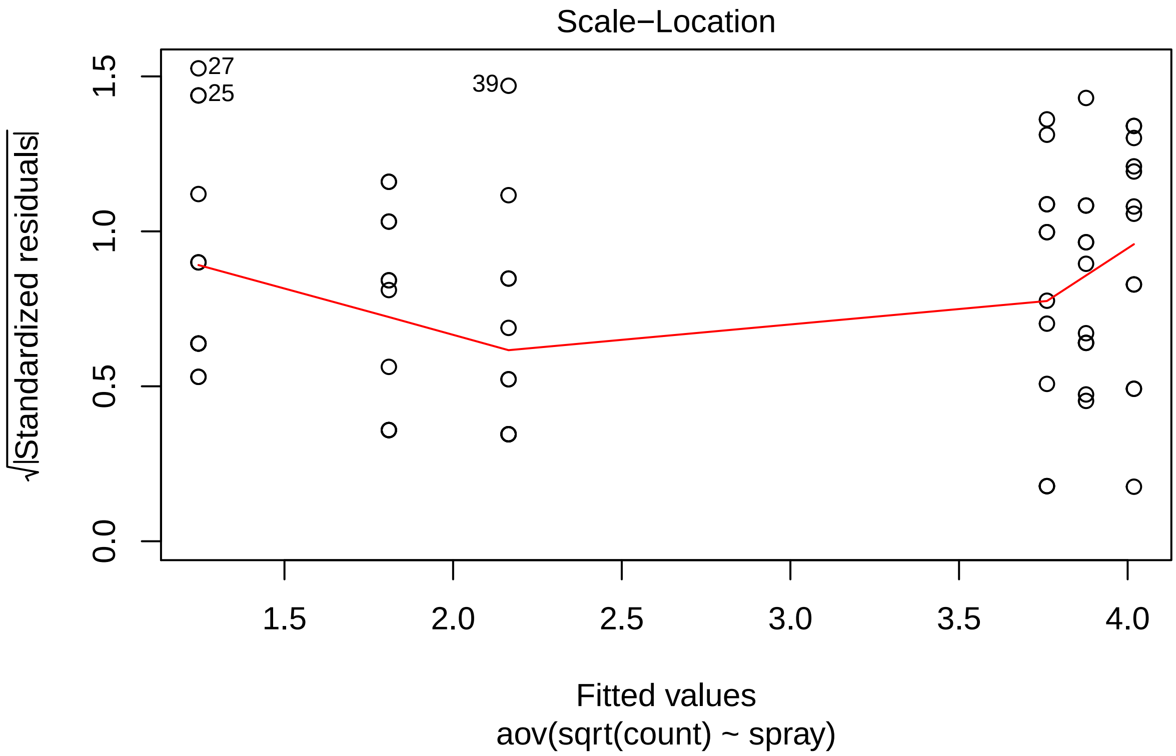
```
Third_model <- aov(sqrt(count) ~ spray, data = InsectSprays)
summary(Third_model)
```

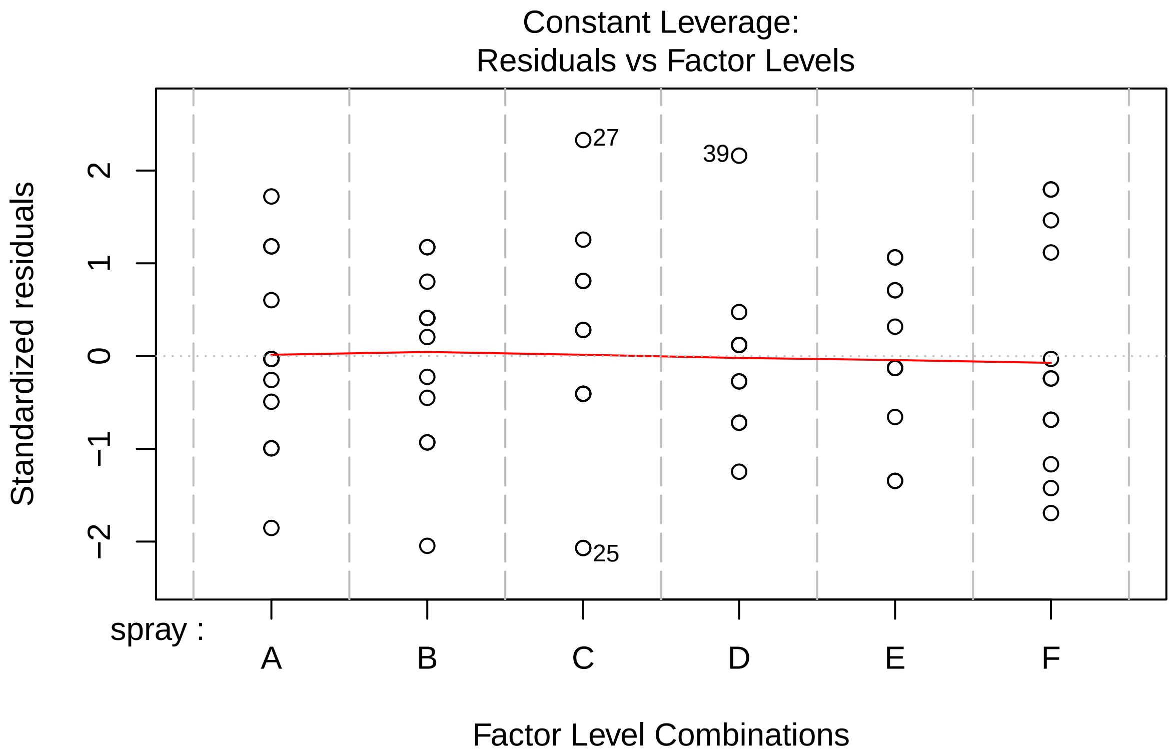
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## spray        5  88.44  17.688    44.8 <2e-16 ***
## Residuals   66  26.06   0.395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(Third_model)
```


Residuals vs Fitted
Residuals
1.5
1.0
0.5
−0.5
−1.5
27
39
25
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
aov(sqrt(count) ~ spray)


Normal Q−Q
Standardized residuals
−2
−1
0
1
2
25
39
27
−2
−1
0
1
2
Theoretical Quantiles
aov(sqrt(count) ~ spray)


Scale−Location
27
25
39
√|Standardized residuals|
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
aov(sqrt(count) ~ spray)


Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
2
1
0
−1
−2
27
39
25
spray :
A
B
C
D
E
F
Factor Level Combinations