Applied Statistics and Data Analysis
Analysis of Variance II

Joaquin Ortega

Fall 2020

Example

A Worked-Out Example

This example is from M.J. Crawley *The R Book*, J. Wiley 2013.

We have an experiment in which crop yields per unit area were measured from 10 randomly selected fields on each of three soil types.

All fields were sown with the same variety of seed and provided with the same fertilizer and pest control inputs.

The question is whether soil type significantly affects crop yield, and if so, to what extent.

A Worked-Out Example

```
results <- read.table('yields.txt',header=T)
attach(results)
str(results)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ sand: int  6 10 8 6 14 17 9 11 7 11
##  $ clay: int  17 15 3 11 14 12 12 8 10 13
##  $ loam: int  13 16 9 12 15 16 17 13 18 14
```

```
head(results, n=4)
```

```
##   sand clay loam
## 1    6   17   13
## 2   10   15   16
## 3    8    3    9
## 4    6   11   12
```

A Worked-Out Example

The function `apply` is used to calculate the mean yields for the three soils

```
apply(results,2,mean)
```

```
## sand clay loam
##  9.9 11.5 14.3
```

Mean yield was highest on loam (14.3) and lowest on sand (9.9).

The question is whether these differences could be due to chance alone

A Worked-Out Example

It will be useful to have all of the yield data in a single vector called `y`.

To create a two-column data frame from a spreadsheet-like `results` where the values of the response are in multiple columns, we use the function called `stack`:

```
frame <- stack(results)
str(frame)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ values: int  6 10 8 6 14 17 9 11 7 11 ...
##  $ ind   : Factor w/ 3 levels "sand","clay",..: 1 1 1 1
```

A Worked-Out Example

You can see that the `stack` function has invented names for the response variable (`values`) and the explanatory variable (`ind`). We change these:

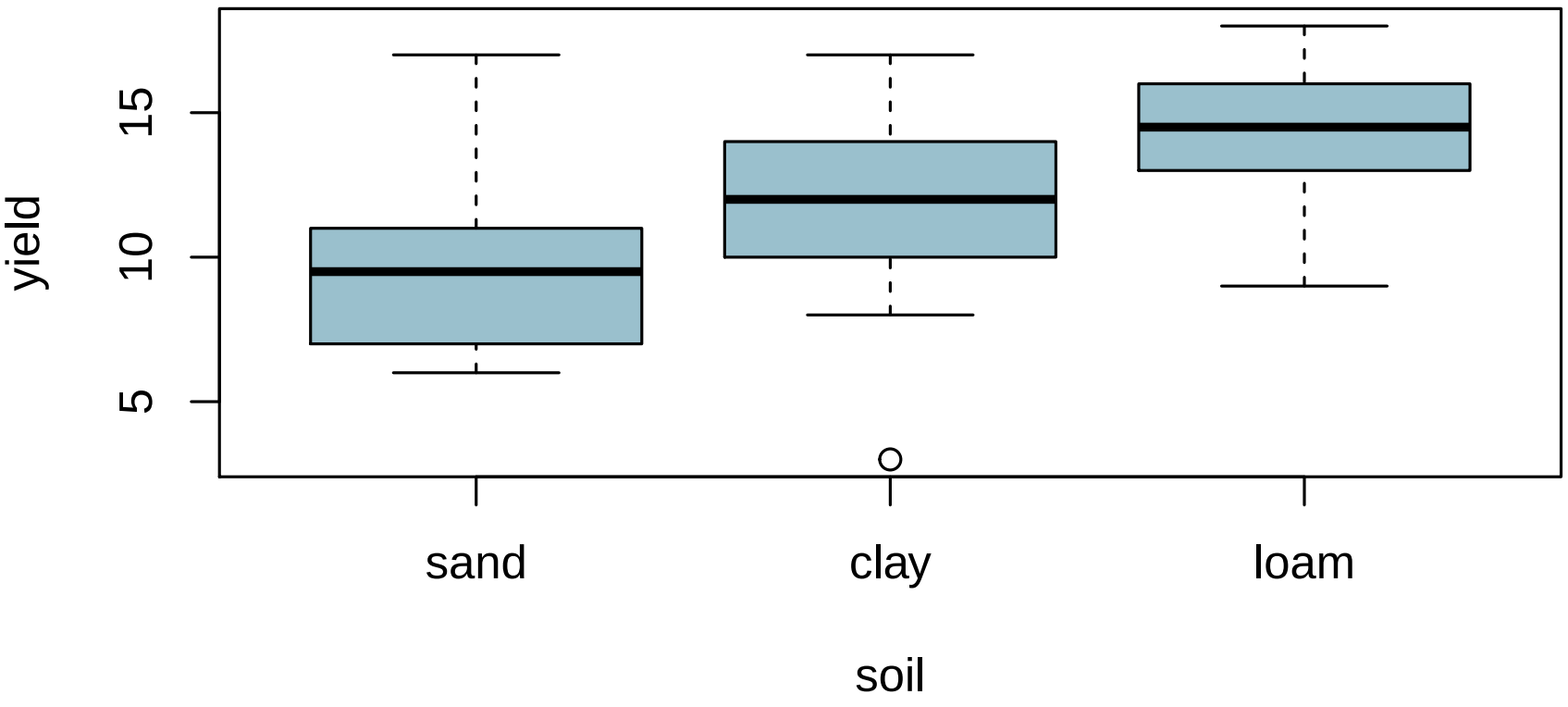
```
names(frame) <- c('yield','soil')
attach(frame)
head(frame)
```

```
##   yield soil
## 1     6 sand
## 2    10 sand
## 3     8 sand
## 4     6 sand
## 5    14 sand
## 6    17 sand
```

A Worked-Out Example

Because the explanatory variable is categorical (three levels of soil type), initial data inspection involves a box plot of `y` against soil:

```
plot(yield~soil,col='lightblue3')
```



A Worked-Out Example

Median yield is lowest on sand and highest on loam Still, there is considerable variation from replicate to replicate within each soil type.

It looks as if yield on loam will turn out to be significantly higher than on sand (their boxes do not overlap), but it is not clear whether yield on clay is significantly greater than on sand or significantly lower than on loam.

Analysis of variance may be used to answer these questions.

A Worked-Out Example

The analysis of variance involves calculating the total variation in the response variable (`yield` in this case) and partitioning it ('analyzing it') into informative components.

In the simplest case, we partition the total variation into just two components, explained variation and unexplained variation.

Explained variation is called the treatment sum of squares (SSA), and unexplained variation is called the error sum of squares (SSE, also known as the residual sum of squares), as defined earlier.

A Worked-Out Example

Let us work through the numbers in R.

From the formula for SST, we can obtain the total sum of squares by finding the differences between the data and the overall mean:

```
sum((yield-mean(yield))^2)
```

```
## [1] 414.7
```

The unexplained variation, SSE, is calculated from the differences between the yields and the mean yields for that soil type:

A Worked-Out Example

```
sand-mean(sand)
```

```
##  [1] -3.9  0.1 -1.9 -3.9  4.1  7.1 -0.9  1.1 -2.9
## [10]  1.1
```

```
clay-mean(clay)
```

```
##  [1]  5.5  3.5 -8.5 -0.5  2.5  0.5  0.5 -3.5 -1.5
## [10]  1.5
```

```
loam-mean(loam)
```

```
##  [1] -1.3  1.7 -5.3 -2.3  0.7  1.7  2.7 -1.3  3.7
## [10] -0.3
```

A Worked-Out Example

We need the sums of the squares of these differences:

```
(sums.vec <- c(sum((sand-mean(sand))^2),
               sum((clay-mean(clay))^2),
               sum((loam-mean(loam))^2)))
```

```
## [1] 112.9 138.5  64.1
```

The total sum of the three terms is

```
sum(sums.vec)
```

```
## [1] 315.5
```

A Worked-Out Example

To get the sum of these totals across all soil types in one single step, we can use `apply` like this:

```
sum(apply(results,2,
            function (x) sum((x-mean(x))^2) ))
```

```
## [1] 315.5
```

Thus *SSE*, the unexplained (or residual, or error) sum of squares, is 315.5.

The extent to which *SSE* is less than *SST* is a reflection of the magnitude of the differences between the means.

A Worked-Out Example

The greater the difference between the mean yields on the different soil types, the greater will be the difference between SSE and SST.

The treatment sum of squares, SSA, is the amount of the variation in yield that is explained by differences between the treatment means.

In this example,

$$SSA = SST - SSE = 414.7 - 315.5 = 99.2$$

A Worked-Out Example

Now we can draw up the ANOVA table.

There are six columns indicating, from left to right,

- the source of variation,
- the sum of squares attributable to that source,
- the degrees of freedom for that source,
- the variance for that source (traditionally called the mean square rather than the variance),
- the F ratio (testing the null hypothesis that this source of variation is not significantly different from zero) and
- the p-value associated with that F value.

We can fill in the sums of squares just calculated, then think about the degrees of freedom:

A Worked-Out Example

Table 2: Anova table for example 2.

Source	Sum of squares	Dof	Mean square	F-ratio	p-value
Soil type	99.2				
Error	315.5				
Total	414.7				

A Worked-Out Example

There are 30 data points in all, so the total degrees of freedom are $30 - 1 = 29$. We lose 1 d.f. because in calculating SST we had to estimate one parameter from the data in advance, namely the overall mean, $\bar{y}_{\bullet\bullet}$, before we could calculate $SST = \sum(y_{ij} - \bar{y}_{\bullet\bullet})^2$.

Each soil type has $n = 10$ replications, so each soil type has $10 - 1 = 9$ d.f. for error, because we estimated one parameter from the data for each soil type, namely the treatment means $\bar{y}_{i\bullet}$ in calculating SSE.

Overall, therefore, the error has $3 \times 9 = 27$ d.f. There were three soil types, so there are $3 - 1 = 2$ d.f. for soil type.

The mean squares are obtained simply by dividing each sum of squares by its respective degrees of freedom (in the same row).

A Worked-Out Example

Table 2: Anova table for example 2.

Source	Sum of squares	Dof	Mean square	F-ratio	p-value
Soil type	99.2	2	$MSA = 49.6$		
Error	315.5	27	$MSE = 11.685$		
Total	414.7	29			

A Worked-Out Example

The error variance, $\hat{\sigma}^2$, is the residual mean square (the mean square for the unexplained variation); this is sometimes called the 'pooled error variance' because it is calculated across all the treatments.

The alternative would be to have three separate variances, one for each treatment:

```
apply(results, 2, var)
```

```
##      sand      clay      loam
## 12.544444 15.388889  7.122222
```

```
mean(apply(results, 2, var))
```

```
## [1] 11.68519
```

A Worked-Out Example

You will see that the pooled error variance $\hat{\sigma}^2 = MSE = 11.685$ is simply the mean of the three separate variances, because (in this case) there is equal replication in each soil type ($n = 10$).

By tradition, we do not calculate the total mean square, so the bottom cell of the fourth column of the ANOVA table is empty.

The F ratio is the treatment variance divided by the error variance, testing the null hypothesis that the treatment means are not significantly different.

If we reject this null hypothesis, we accept the alternative hypothesis that at least one of the means is significantly different from the others.

A Worked-Out Example

Table 2: Anova table for example 2.

Source	Sum of squares	Dof	Mean square	F-ratio	p-value
Soil type	99.2	2	$MSA = 49.6$	4.24	
Error	315.5	27	$MSE = 11.685$		
Total	414.7	29			

A Worked-Out Example

The question naturally arises at this point as to whether 4.24 is a big number or not.

If it is a big number, then we reject the null hypothesis. If it is not a big number, then we do not reject the null hypothesis.

We calculate the p value associated with our test statistic of 4.24 using the function `pf` for cumulative probabilities of the F distribution like this:

```
1-pf(4.24,2,27)
```

```
## [1] 0.02503987
```

A Worked-Out Example

Table 2: Anova table for example 2.

Source	Sum of squares	Dof	Mean square	F-ratio	p-value
Soil type	99.2	2	$MSA = 49.6$	4.24	0.025
Error	315.5	27	$MSE = 11.685$		
Total	414.7	29			

A Worked-Out Example

That was a lot of work. R can do the whole thing in a single line:

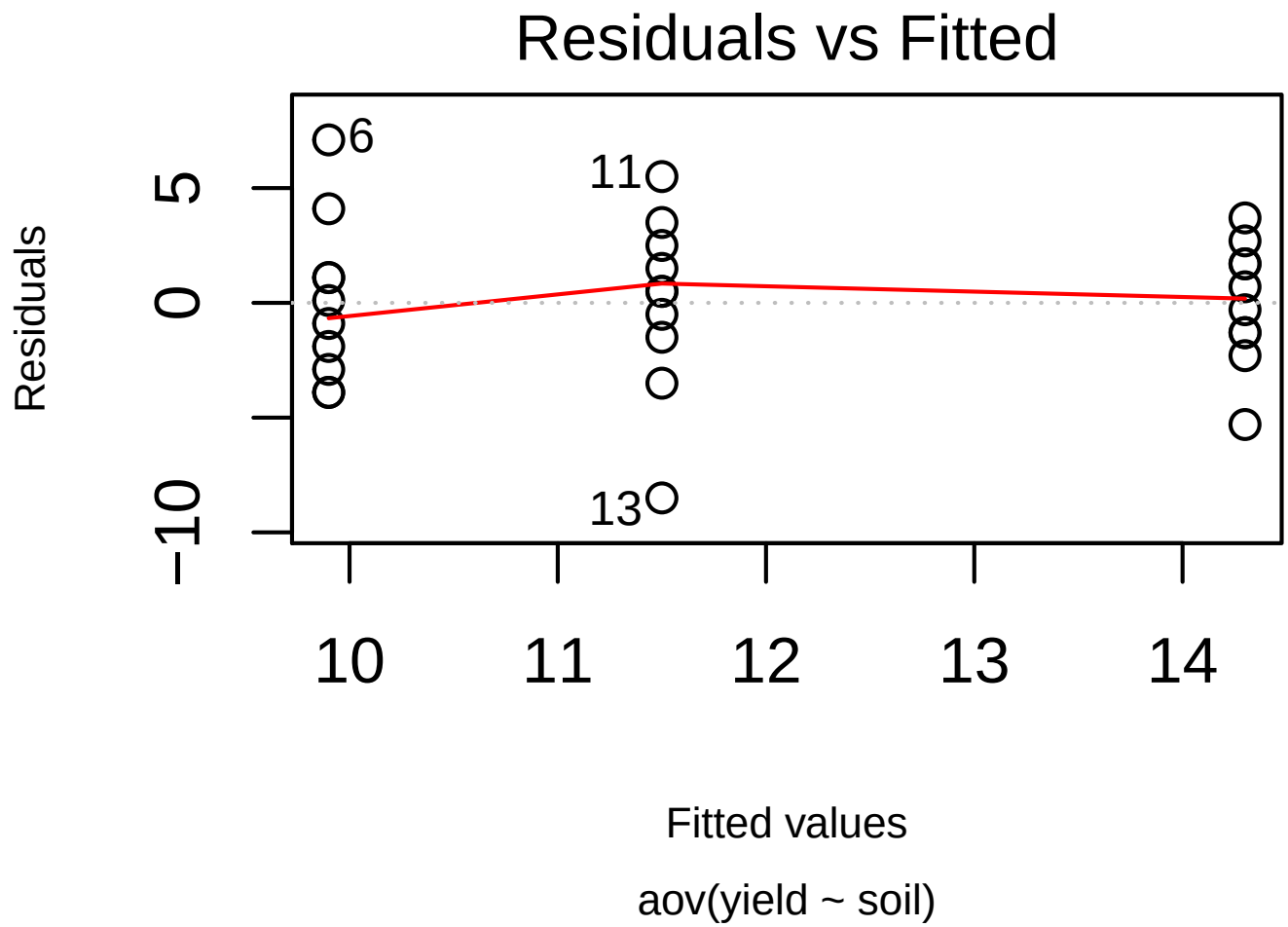
```
summary(aov(yield~soil))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## soil         2   99.2   49.60   4.245  0.025 *
## Residuals   27  315.5   11.69
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A Worked-Out Example

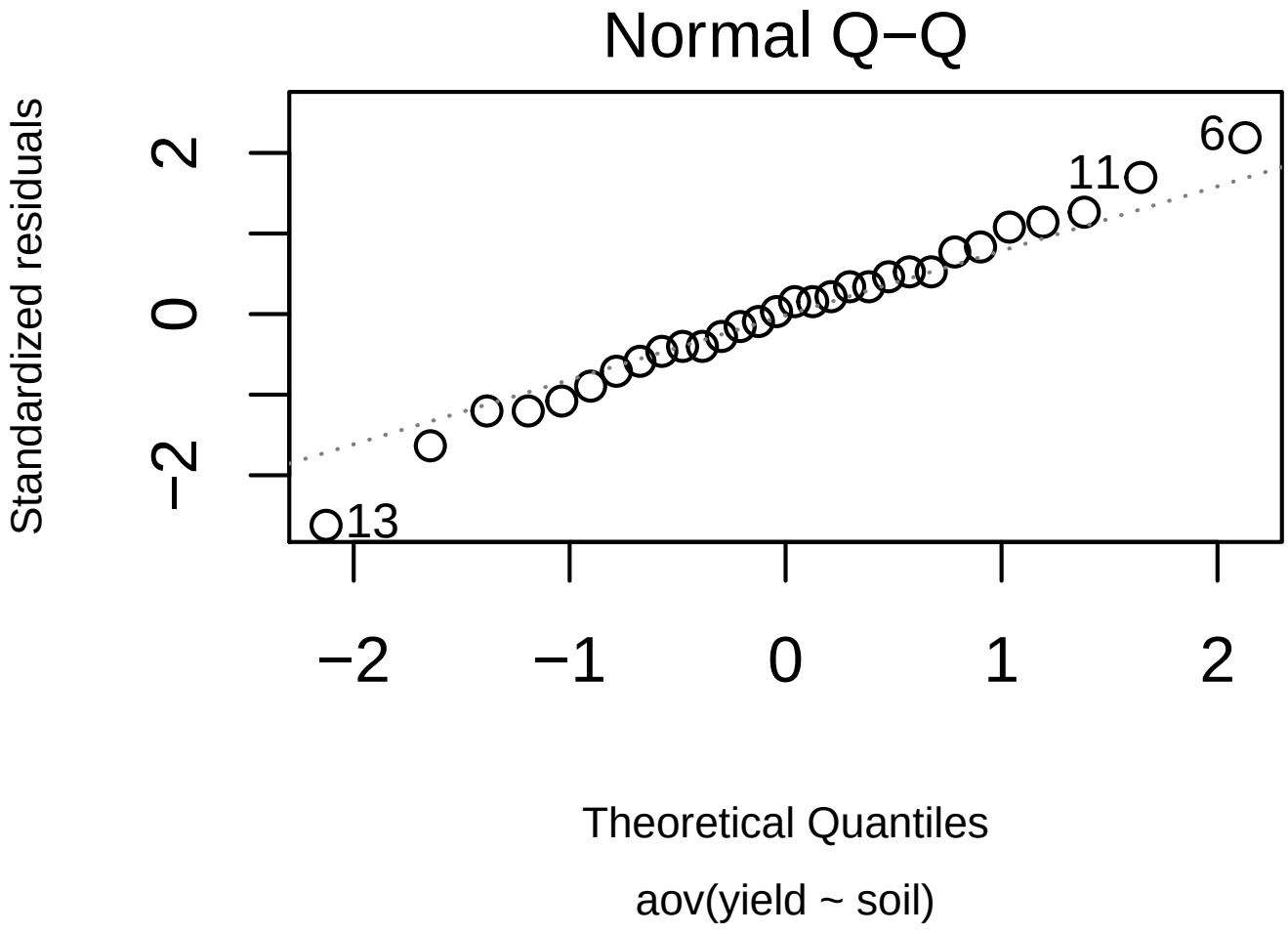
The next thing we would do is to check the assumptions of the `aov` model. This is done using `plot`.

```
plot(aov(yield~soil), which = 1,
     cex.lab=0.7, cex.sub=0.7)
```



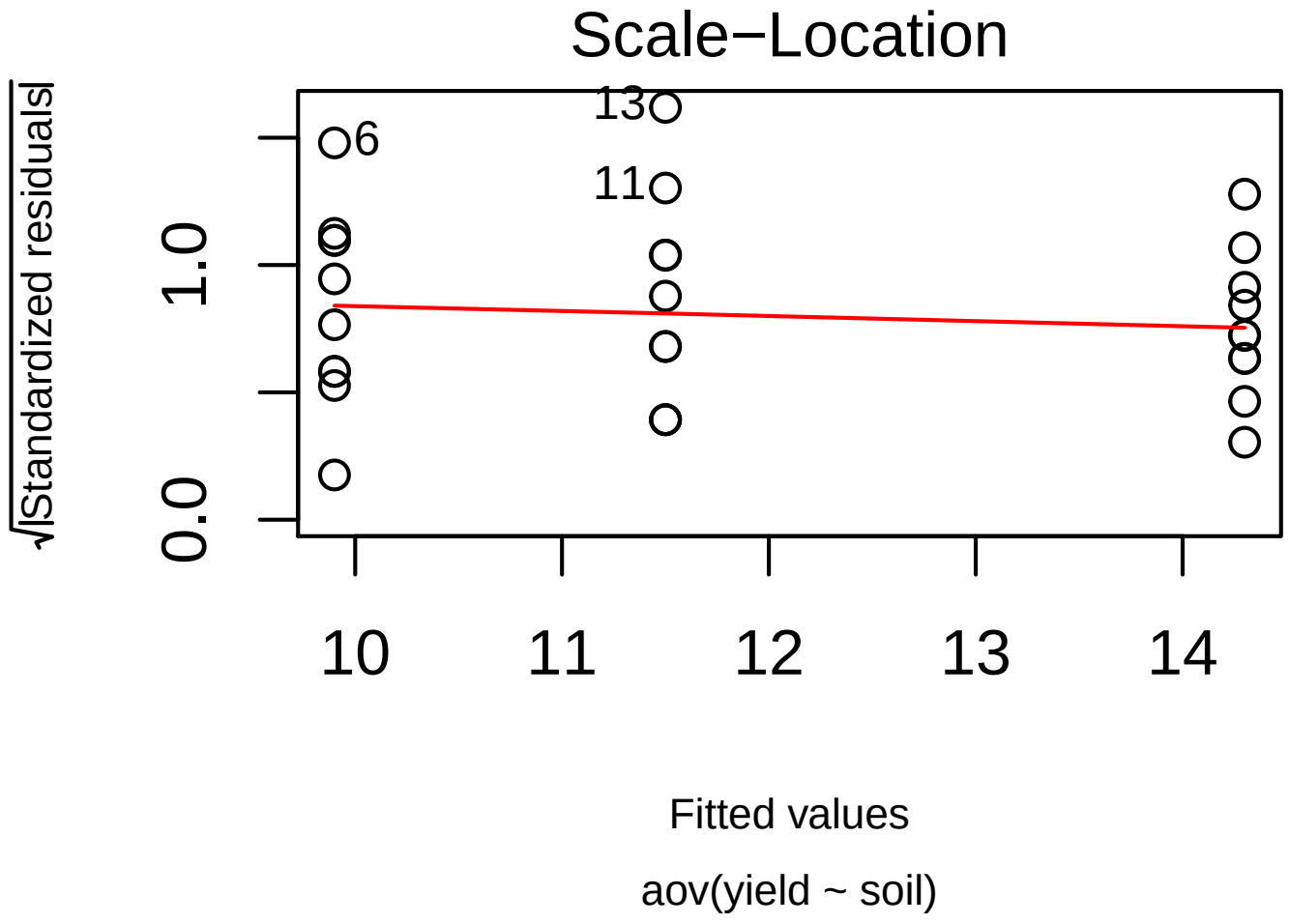
A Worked-Out Example

```
plot(aov(yield~soil), which = 2,
     cex.lab=0.7, cex.sub=0.7)
```



A Worked-Out Example

```
plot(aov(yield~soil), which = 3,
     cex.lab=0.7, cex.sub=0.7)
```



A Worked-Out Example

```
plot(aov(yield~soil), which = 4,
     cex.lab=0.7, cex.sub=0.7)
```