# One-way ANOVA with three levels

## Structure of the model

An ANOVA with three levels can be written with three $\beta$ parameters:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

where:

- $Y$ is a vector of observed values
- $\beta_1$ is a parameter representing the mean at level 1, treated as unknown but constant.
- $X_1$ is a binary predictor that is 1 if this $y$ has the factor at level 1, zero otherwise
- $\beta_2$ is a parameter representing the mean at level 2, treated as unknown but constant.
- $X_2$ is a binary predictor that is 1 if this $y$ has the factor at level 2, zero otherwise
- $\beta_3$ is a parameter representing the mean at level 3, treated as unknown but constant.
- $X_3$ is a binary predictor that is 1 if this $y$ has the factor at level 3, zero otherwise
- $e$ is a vector of independent, identically distributed $N(0, \sigma_e)$ random variables
- $\sigma_e$ is the standard deviation of the error or residual terms $e$, which are assumed to have mean zero.

## Example: Sepal width by species

Read the data:

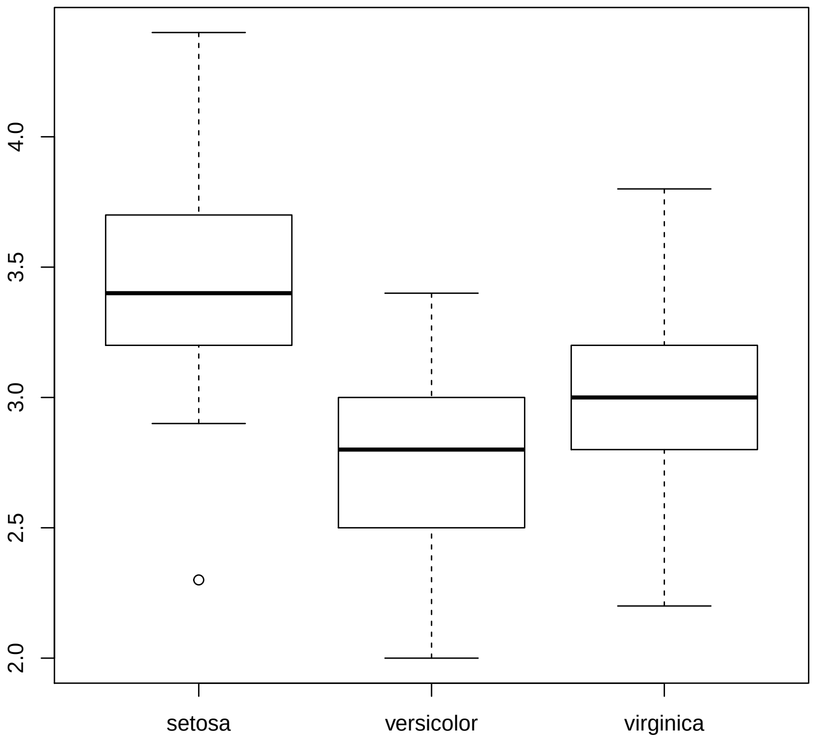
```
data(iris)

Sepal.Width   = iris$Sepal.Width      #get sepal width

Species       = iris$Species
```

Check out the data with a boxplot

```
boxplot(Sepal.Width~Species)
```



Now run the one-way ANOVA.

As long as Species is a factor, R will determine the number of levels.

```
aov1 = aov(Sepal.Width ~ Species)

summary(aov1)

##                Df Sum Sq Mean Sq F value Pr(>F)
## Species         2  11.35   5.672   49.16 <2e-16 ***
## Residuals     147  16.96   0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is the standard ANOVA table.

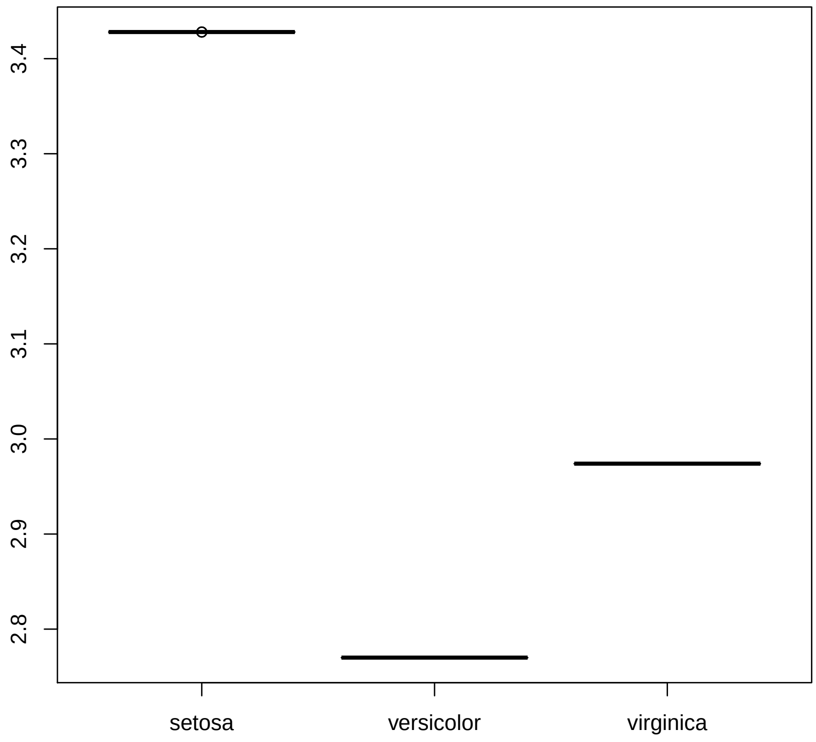
It has one line for each factor, and one line for `Residuals`.

Since we have only one factor, it has only two lines.

The `Species` line is for the species factor. The F-statistic has `Pr(>F)` much less than .05, so it is very unlikely that this data was generated from a dataset where the mean sepal width values of the three species are nearly the same.

Like any linear model, this one has fitted values. We can examine them with a boxplot:

```
boxplot(aov1$fitted.values~Species)
```



The fitted values in any one-way ANOVA are the means for the levels:

Compute the means of Sepal.Width by level of Species:

```
tapply(Sepal.Width,Species,mean)

##     setosa versicolor  virginica
##      3.428      2.770      2.974
```

These can be obtained from the coefficients, though the presentation can be confusing:

```
aov1$coefficients

##       (Intercept) Speciesversicolor  Speciesvirginica
##             3.428            -0.658            -0.454
```

You would interpret this to say that the expected sepal width for setosa is the `(Intercept)` value.

To obtain the expected sepal width for versicolor, add the `(Intercept)` and `Speciesversicolor` values.

For virginica, add the `(Intercept)` and `Speciesvirginica`

This will produce the same numbers as we got by computing the mean sepal width values by species.
Since we have three levels, we need to determine which differences are significant.

We will use the Tukey honest significant difference multiple comparison test.

```
TukeyHSD(aov1)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sepal.Width ~ Species)
##
## $Species
##                        diff         lwr        upr     p adj
## versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
## virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
## virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

Differences that are significant have the same sign for `lwr` and `upr`. The approximate significance level is in the `p adj` column.

In this case, each pair of means is significantly different.