Processing and Analysis of Biological Data

Chapter 3. The Linear Model II: Analysis of Variance

Øystein H. Opedal, Department of Biology, Lund University

2024-11-10

Analysis of variance (ANOVA)

When our predictor variables are categorical (factors), linear models are used to perform analyses of variance. The parameter estimation works in much the same way as in regression, except that instead of estimating regression slopes, we are estimating group effects.

The aim of our ANOVA analysis is to partition the variance in our response variable into a set of additive components (recall, variances are additive). Based on this, we can evaluate whether the variance among groups is greater than the variance within groups, or more so than expected by chance.

The ANOVA framework is based on calculating "sums of squares" (SS), the sum of the squared deviations of each group mean from the grand mean (the variance among groups), and of each datapoint from either the grand mean (which is equal to the total variance in the data), or the group means (the residual or within-group variance).

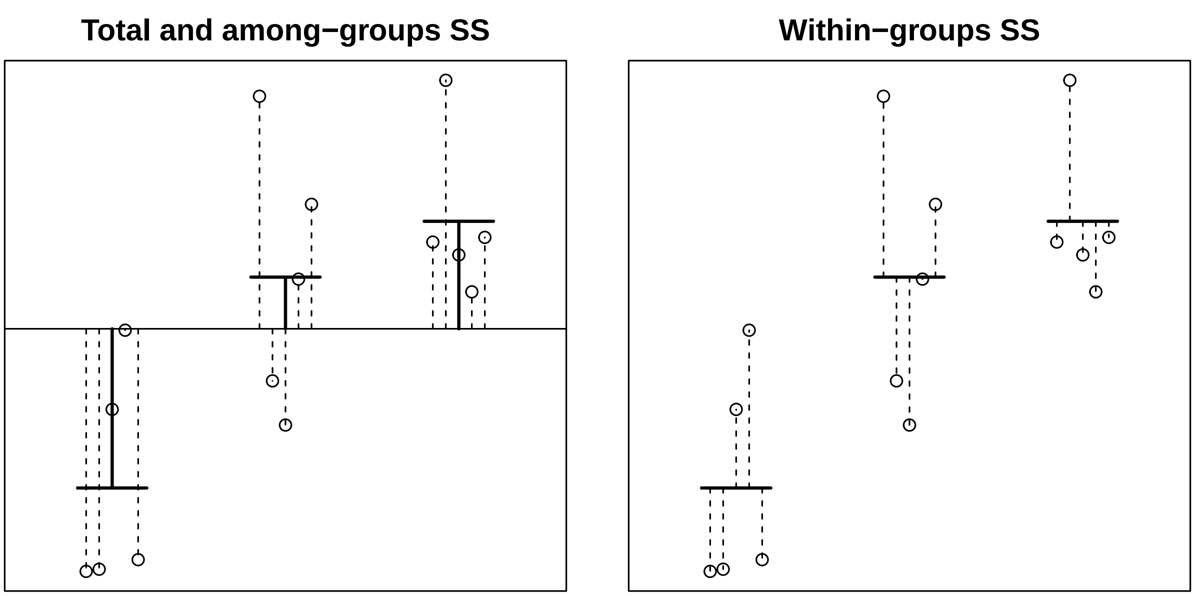
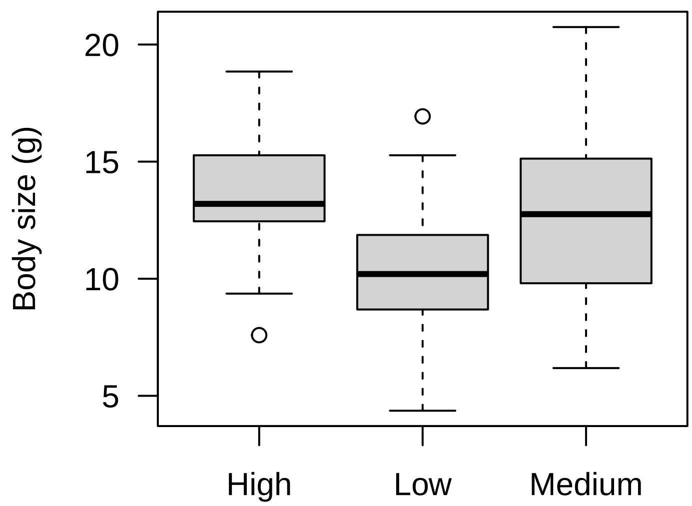


Figure 1: Illustration of total, among-groups, and within-groups sums of squares. The total sum of squares is the total variance in the data, which can be partitioned into among-group and within-group (residual) components

To understand how the ANOVA analysis works, let's simulate some data, fit a linear model, and perform an ANOVA.

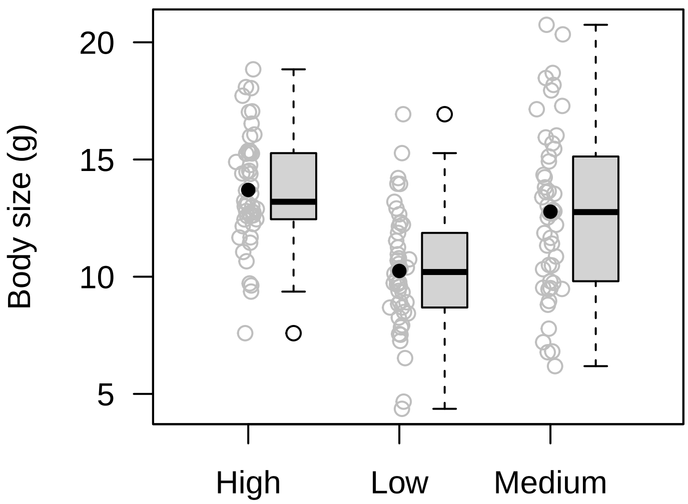
```
set.seed(100)
groups = as.factor(rep(c("Low", "Medium", "High"), each=50))
x = c(rnorm(50, 10, 3), rnorm(50, 13, 3), rnorm(50, 14, 3))

plot(groups, x, las=1, xlab="", ylab="Body size (g)")
```



Plots like this are called "boxplots", and can be useful for visualising the distribution of data across factor levels or other groups. With the default settings, the boxes range from the 1st to the 3rd quartile, i.e. they span 50% of the data. The thick lines show the median, the "whiskers" extend to 1.5 times the inter-quantile range, and individual circles show outliers. This representation allows us to assess whether the data are roughly normally distributed within each group, which will tend to yield normally distributed residuals within each group, an assumption of the ANOVA model.

Boxplots are not directly useful for discussing ANOVA and variance partitioning though, and plots that show all the data can be more informative. Here is a rather elaborate plot combining a scatterplot (with values slightly "jittered" along the x-axis for clarity) and a boxplot.

EXTRA EXERCISE: Reproduce a plot similar to this for the ANOVA exercise.



Just as for linear regression, we fit the model with the `lm` function and save the results to an object (here `m`). To perform an ANOVA based on the fitted model, we call the `anova` function.

```
m = lm(x~groups)
anova(m)
```

```
## Analysis of Variance Table
##
## Response: x
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## groups      2  319.97 159.985  19.591 2.866e-08 ***
## Residuals 147 1200.43   8.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table contains a lot of information. First, we learn about the number of degrees of freedom for each variable. For the `groups` variable (our focal factor), the 2 degrees of freedom is the number of groups in our data (3) - 1. The minus 1 comes from the fact that we had to estimate the mean in the data to obtain our sums of squares (the sum of the square deviations of data points from their group means). Similarly for residual degrees of freedom, we have 150 - 2 - 1, where the 2 comes from estimating the two contrasts (difference of group 2 and 3 from group 1), and the 1 is still the estimated mean.

The `Sum Sq` are the sums of squares, i.e. the sum of the squared deviations of each observation from the grand mean. The total sum of squares (SS_T) divided by $n - 1$ gives the total variance of the sample.

```
SS_T = 319.97+1200.43
SS_T/(150-1)
```

```
## [1] 10.20403
```

```r
var(x)
```

```
## [1] 10.20403
```

We can easily get the proportion of variance explained by the `groups` variable, which is the same as the r^2 for the model.

```r
319.97/SS_T
```

```
## [1] 0.2104512
```

The `Mean Sq` is the variance attributable to each variable, conventionally called the mean sum of squares (the sum of squares divided by the degrees of freedom). The F-ratio (the test statistic in an ANOVA) is computed as the mean sum of squares for the focal variable divided by the mean residual sum of squares. Thus, it represents the ratio of the among-group variance to the within-group variance, but also the sample size which gives the residual degrees of freedom and thus, all else being equal, larger sample gives lower mean sum of squares and thus higher F-ratios and lower P-values.

In an ANOVA, a statistically supported among-group variance component such as the one above indicates that at least one group mean is different from the others. To further assess which groups are different, we can extract the typical summary table of the linear model.

```r
summary(m)
```

```
##
## Call:
## lm(formula = x ~ groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5887 -1.5596 -0.0987  1.6274  7.9729
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.7006     0.4041  33.901  < 2e-16 ***
## groupsLow     -3.4561     0.5715  -6.047 1.16e-08 ***
## groupsMedium  -0.9277     0.5715  -1.623    0.107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.858 on 147 degrees of freedom
## Multiple R-squared:  0.2105, Adjusted R-squared:  0.1997
## F-statistic: 19.59 on 2 and 147 DF,  p-value: 2.866e-08
```

This contains some of the same information as the ANOVA table, but we now also obtain parameter estimates. The first parameter, the intercept, corresponds to the estimated mean for the first level of the `groups` factor. In this example this happens to be 'High', because H comes before L and M in the alphabet. The next two estimates represents *contrasts* from the reference group, and the associated hypothesis tests test the null hypothesis that the group has the same mean as the reference group.

The summary table also gives us directly the r^2, which is a simple ANOVA is defined as $1 - SS_E/SS_T$, where SS_E is the sum of squares for the error term (residuals), and SS_T is the total sum of squares. (Control question: why could we do it even more simply above?).

The parameter estimates allow us to quantify the effect size, i.e. the magnitude of the difference between the groups. A useful way to report such differences is to compute the % difference (the contrast divided by the mean of the reference group, here -3.456/13.701 = -0.252), so that we can say that "Individuals in the low-food treatment were 25.2% smaller than individuals in the high-food treatment".

Note that if we want a different reference group, we can change the order of the factor levels.

```
groups = factor(groups, levels=c("Low", "Medium", "High"))
m = lm(x~groups)
summary(m)
```

```
##
## Call:
## lm(formula = x ~ groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5887 -1.5596 -0.0987  1.6274  7.9729
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.2445     0.4041  25.349  < 2e-16 ***
## groupsMedium   2.5284     0.5715   4.424 1.88e-05 ***
## groupsHigh     3.4561     0.5715   6.047 1.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.858 on 147 degrees of freedom
## Multiple R-squared:  0.2105, Adjusted R-squared:  0.1997
## F-statistic: 19.59 on 2 and 147 DF,  p-value: 2.866e-08
```

Sometimes we also want to suppress the intercept of the model, and thus estimate the mean and standard error for each level of the predictor. We can do this by adding `-1` to the model formula (what comes after the ~ sign). This could be useful for example, if we wanted to obtain the estimated mean for each group, associated for example with a 95% confidence interval.

```
m = lm(x~groups-1)
summary(m)$coef
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## groupsLow    10.24453  0.4041333 25.34937 1.586298e-55
## groupsMedium 12.77295  0.4041333 31.60578 1.980446e-67
## groupsHigh   13.70064  0.4041333 33.90129 2.276884e-71
```

```
confint(m)
```

```
##                  2.5 %   97.5 %
## groupsLow     9.445865 11.04319
## groupsMedium 11.974287 13.57161
## groupsHigh   12.901979 14.49930
```

Two-way ANOVA

Analyses of variance can also be performed with more than one factor variable. If we have two factors, we can talk about two-way ANOVA, and so on. A typical example from biology is when we have performed a factorial experiment, and want to assess the effects of each experimental factor and their potential interaction.

With two factors, a full model can be formulated as `y ~ factor1 * factor2`. Recall that in `R` syntax, the * means both main effects and their interaction, while a : means only the interaction. A detectable interaction term in this model would indicate that the effect of factor 1 depends on the level of factor 2 (and *vice versa*). If we are analysing an experiment where we have manipulated both temperature and nitrogen supply, an interaction would mean that the effect of temperature depend on the nitrogen level.

Data exercise: analysing a factorial experiment

The following data are from an experiment where female butterflies reared on two different host plants were allowed to oviposit on the same two host plants. The data include developmental time for the larvae, the adult weight, and the growth rate of the larvae.

Analyse the data to assess effects of maternal vs. larval host plant on one or more response variable. Interpret the results and produce a nice plot to illustrate.

```
dat = read.csv("datasets/butterflies.csv")
names(dat)
```

```
## [1] "LarvalID"        "LarvalHost"      "Sex"             "MaternalHost"
## [5] "MotherID"        "DevelopmentTime" "AdultWeight"     "GrowthRate"
```

As a first step, let us compute some summary statistics, like the mean development time for each combination of larval and maternal host plant.

```
dat$MaternalHost = paste0(dat$MaternalHost, "M")
dat$LarvalHost = paste0(dat$LarvalHost, "L")
means = tapply(dat$DevelopmentTime, list(dat$MaternalHost, dat$LarvalHost), mean)
means
```

```
##           BarbareaL BerteroaL
## BarbareaM  21.69608  27.00000
## BerteroaM  23.51282  31.01923
```

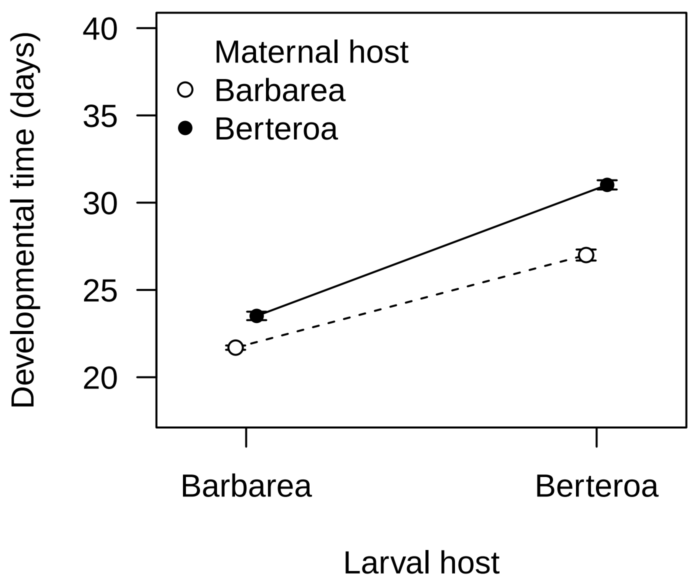


Figure 2: Larval developmental time depending on larval and maternal host plant. Error bars indicate standard errors