ANALYSIS OF VARIANCE

Master of Statistics

Lesson 1

Dr. Francesca Solmi

Contents

1 Introduction

1.1 ANOVA problem

ANOVA problem

This course will deal with ANalysis Of VAriance (ANOVA) problems. We will learn how to analyze single- and multi-factor (experimental or observational) studies. Hence we will focus on the problem of *comparing* different (more than 2) populations, in terms of some outcome of interest. The framework we are working in is characterized by the presence of:

- a quantitative outcome (response variable) Y of interest,
- one or more fixed factors (denoting a medical treatment, a machine, a characteristic of a population, ...), characterized by two or more levels each.

In practice, after the study is designed, we observe realizations of the response variable, $\boldsymbol{y} = \{y\}_{ij}$, for the j^{th} replicate of the i^{th} level of the factor(s) (treatment), with $j = 1, \ldots, n_i$, $i = 1, \ldots, r$ and $\sum_{i=1}^{r} n_i = n_T$. Attention to the notation!

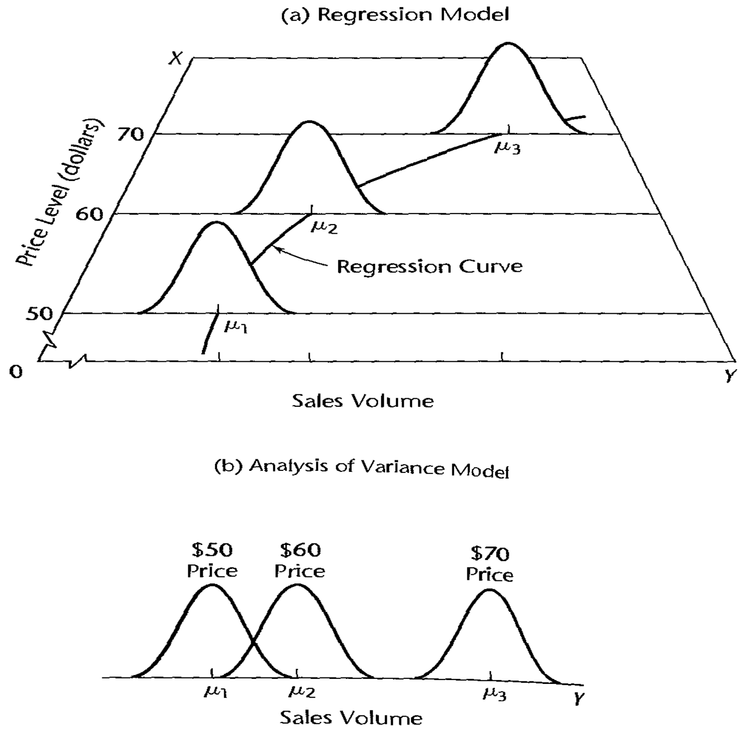
1.2 Relation with linear regression

A strong relation

There is a strong relation between *linear regression* and *ANOVA* models. In both cases we want to model the distribution of the response Y using some explanatory variables. The difference is in the nature of these explanatory variables: if they are categorical (or categorized), then the ANOVA problem can be solved by fitting a linear regression model which uses particular explanatory variables (e.g. dummy variables).

Different dimension of the problem

Pricing study example:



Different model formulations

Look at the indexes and at the covariates:

(a) Regression Model $\big(y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \epsilon_i\big)$

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ \cdots \\ \\ \cdots \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ 1 & z_n & z_n^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \cdots \\ \\ \cdots \\ \cdots \\ \epsilon_n \end{pmatrix}$$

(b) Analysis of Variance Model $(y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \epsilon_{ij})$

$$\begin{pmatrix} y_{1\,1} \\ y_{1\,2} \\ \cdots \\ \cdots \\ \\ \cdots \\ \cdots \\ y_{3\,n_3} \end{pmatrix} = \begin{pmatrix} 1 & x_{1\,1,1} & x_{1\,1,2} \\ 1 & x_{1\,2,1} & x_{1\,2,2} \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ 1 & x_{3\,n_3,1} & x_{3\,n_3,2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{1\,1} \\ \epsilon_{1\,2} \\ \cdots \\ \cdots \\ \\ \cdots \\ \cdots \\ \epsilon_{3\,n_3} \end{pmatrix}$$

Different use of the covariates

(a) If we consider the price level as a continuous variable, z, the linear model can fit, for instance, a quadratic function of z to explain the mean response.

(b) If we consider the price level as a factor with $r = 3$ levels, the ANOVA model can use, for instance, $(r - 1) = 2$ indicator (dummy) variables as predictors:

$$x_{1\,i} = \begin{cases} 1 & \text{if price level } (z_i) \ = 60 \\ 0 & \text{otherwise} \end{cases}$$

$$x_{2\,i} = \begin{cases} 1 & \text{if price level } (z_i) \ = 70 \\ 0 & \text{otherwise} \end{cases}$$

1.3 Structure of the course

Structure

The course will be structured as follows:

- Most of the concepts discussed will refer to chapters in the book *Applied Linear Statistical Models, 5th Edition, Kutner et al. (2005)*,
- Most lessons will be in **video format**. The students are expected to watch the videos following the **schedule of the course** (schedule and summary available on Blackboard in the "Course Information" folder).
- There are **exercises** related to sets of lessons. The students are expected to solve the exercises and check their results with the solutions, which will be provided as R code.
- There will be **contact moments** in which both theoretical concepts and exercises will be discussed. The students are expected to use these contact moments to ask questions to the teachers.
- There will be two multiple choice **quizzes** on scheduled dates, in which the students will be tested on the concepts thought in the (video) lectures. These quizzes will be evaluated and the score will count up to **4 points** for the final mark.
- There will be a **final written exam**. The exam will count up to **16 points** for the final mark.
- Points division (1^{st} chance exam):

Quizzes	4
Written exam	16
Final mark	20

- In the 2^{nd} chance exam, the evaluation will consist only of the written exam score (0-20 points).
- The key dates for (video) lectures, contact moments, quizzes and exercises are available on Blackboard in the "Course Information" folder.

Course material

- The R code to deal with all taught concepts is provided in the slides/handouts.
- The slides/handouts and the videos for the each lesson are available on Blackboard in the "Course Material" folder, at scheduled dates.
- All the exercises and the R code with the solutions are available on Blackboard in the "Assignments" folder, at scheduled dates.

Course contents

1. Introduction (lesson 1)
2. Single Factor Analysis of Variance (lesson 1)
3. Analysis of Factor Level Effects (lesson 2)
4. Diagnostics and Remedial Measures (lesson 3)
5. Model reformulation and regression approach (lesson 4)
6. Two-way ANOVA (Equal sample size) (lessons 5 and 6)
7. Two-way ANOVA (Unequal sample size) (lesson 7)
8. Estimable Functions (lesson 8)
9. Type I to III sum of squares for Two-way ANOVA (lesson 8)

2 Single Factor Analysis of Variance

2.1 2 samples t test (recall)

Comparing 2 populations

Let us start from the (simpler) problem of comparing 2 populations. This is a *particular case* of our ANOVA problem, where the factor has only 2 levels. In general, the hypotheses we want to test are:

$$\begin{cases} H_0: & \mathcal{F} = \mathcal{G} \\ H_1: & \mathcal{F} \neq \mathcal{G} \end{cases}$$

where $\mathcal{F}$ and $\mathcal{G}$ are the underlying response distributions for the two populations.

The t test: tested hypotheses

When parametric assumptions are made on $\mathcal{F}$ and $\mathcal{G}$, the problem can be solved with the so-called 2 samples t test. After assuming that the observed responses are *independent* realizations from the two populations, which are both *normally distributed* with (just an option) *the same variance* and means respectively μ_1 and μ_2, the hypotheses we want to test become:

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases} .$$

The t test: solution

The problem is solved, as any statistical testing problem, by:

(1) choosing a suitable test statistic T (very important, T is a random variable),

(2) computing T on the observed dataset, getting the value t^{obs},

(3) comparing t^{obs} with the distribution of T under H_0,

(4) rejecting H_0 if and only if $t^{obs} \in R$, where R is the rejection region for the test (at some fixed significance level).

In this case, with our assumptions,

- T is chosen as
$$T = |\bar{Y} - \bar{X}| \, / \left(S\sqrt{1/n_1 + 1/n_2} \right)$$
with $\bar{X} = \sum_{i=1}^{m} X_i/n_1$, $\bar{Y} = \sum_{j=1}^{n} Y_j/n_2$ and $S^2 = \left[\sum_{i=1}^{m} (X_i - \bar{X})^2 + \sum_{j=1}^{n} (Y_j - \bar{Y})^2 \right] /(n_1 + n_2 - 2)$,
- its distribution under H_0 is $T \sim t_{n_T - 2}$,
- the rejection region is given by $t^{obs} > t_{1-\alpha/2;n_T-2}$, where $t_{1-\alpha/2;n_T-2}$ is the $(1-\alpha/2)\%$ quantile of a Student t with $n_T - 2$ degrees of freedom, α is the level of significance and n_T is the total sample size.

2.2 ANOVA: Hypotheses and model definition

Basic ideas

Kutner et al. (2005), Chapter 16.

A general ANOVA problem differs from the 2 samples one in the number of populations that are compared (more than 2). The basic idea remains the same: corresponding to each factor level we have

and certain underlying distribution for the response ($\mathcal{F}_1$, $\mathcal{F}_2$,...,$\mathcal{F}_r$). We want to test the hypotheses:

$$\begin{cases} H_0 : & \mathcal{F}_1 = \mathcal{F}_2 = \cdots = \mathcal{F}_r \\ H_1 : & \text{at least one } \mathcal{F}_i \text{ is different} \end{cases} .$$

Assumptions and hypotheses

Also in this case we can make some assumptions:

- each probability distribution is *normal*,
- each probability distribution has the *same variance*,
- the responses from each factor levels are *independent* realizations from the corresponding probability distributions.

Hence the hypotheses we want to test become:

$$\begin{cases} H_0 : & \mu_1 = \mu_2 = \cdots = \mu_r \\ H_1 : & \text{at least one } \mu_i \text{ is different} \end{cases} .$$

Cell means model

The ANOVA model can be stated, in presence of r factor levels, as

$$y_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, \dots, r$ and $j = 1, \dots, n_i$. Thus y_{ij} are the observed responses for the j^{th} replicate of the i^{th}

level of the factor(s), μ_i, $i = 1, \dots, r$, are the parameters and ϵ_{ij}, $i = 1, \dots, r$ and $j = 1, \dots, n_i$, are the random errors, assumed to be realizations of $N(0, \sigma^2)$. Through some simple calculations we can prove

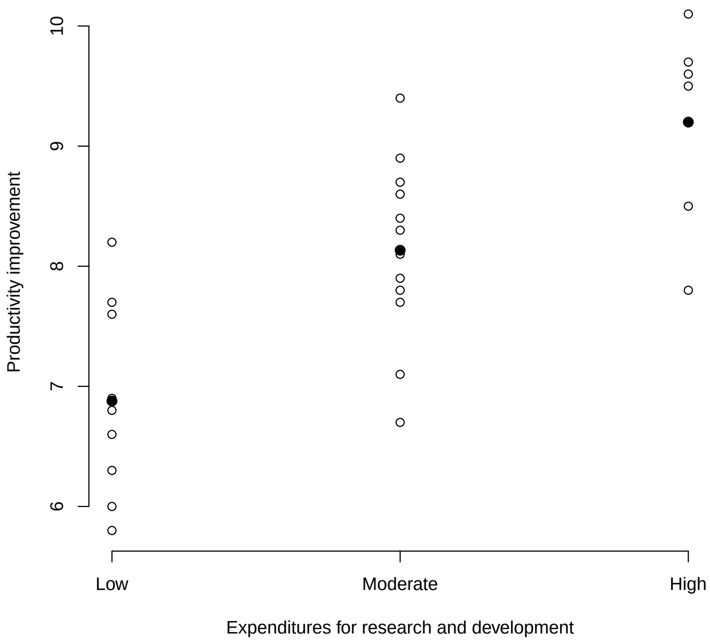
that the assumptions we stated before hold in such model. The y_{ij} are sum of the constant term μ_i and the random error term ϵ_{ij}, which means that (check why!):

- $E[Y_{ij}] = \mu_i$,
- $V[Y_{ij}] = \sigma^2$, *constant*,
- Y_{ij} are *normally distributed*,
- y_{ij} are *independent* realizations from the corresponding distributions.

R example: the data

This is a dataset on productivity improvement from a sample of firms. The firms are classified in terms of their average expenditures in research and development in the past three years (low, moderate, high). Let us get started with R:

```
> ### Importing the data
>
> data = read.table(file.choose(),col.names=c("y","group","j_index"))  # import file data_ch16pr07.t
>
> str(data)
’data.frame’:   27 obs. of  3 variables:
 $ y      : num  7.6 8.2 6.8 5.8 6.9 6.6 6.3 7.7 6 6.7 ...
 $ group  : int  1 1 1 1 1 1 1 1 1 2 ...
 $ j_index: int  1 2 3 4 5 6 7 8 9 1 ...
> dim(data)
[1] 27  3
>

> ### Displaying the data
> attach(data)
> plot(group,y,xlab="Expenditures for research and development",
+              ylab="Productivity improvement",axes=F)
> group_means = c(mean(y[group==1]),mean(y[group==2]),mean(y[group==3]))
> points(c(1,2,3),group_means,pch=19,cex=1.2)
> axis(side=1, at = c(1,2,3), labels = c("Low","Moderate","High"))
> axis(side=2)
```



2.3 LS (ML) estimations and residuals

Estimation methods

The parameters of the ANOVA model, μ_i, $i = 1, \ldots, r$, need to be estimated. Least squares (LS) and the maximum likelihood (ML) estimates can be used (minimum variance unbiased estimators):

- LS: the estimates for μ_i are given by the quantities that minimize the sum of squared deviations of the observations from the expected model.
- ML: the estimates for μ_i are given by the quantities that maximize the (log-)likelihood function calculated on the observed dataset.

As in normal errors regression models, LS and ML estimates coincide for ANOVA models.

Let us consider the LS method: we need to minimize the sum of squared deviations of the observations from their expected values. We know that $E[Y_{ij}] = \mu_i$, $\forall i = 1, \ldots, r$ and $\forall j = 1, \ldots, n_i$. Hence we need to minimize

$$\begin{aligned} Q =& \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \\ =& \sum_{j=1}^{n_i} (y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_i} (y_{2j} - \mu_2)^2 + \cdots + \sum_{j=1}^{n_r} (y_{rj} - \mu_r)^2 , \end{aligned}$$

which coincides with minimizing all the terms of the sum. The solution is then given by $\hat{\mu}_i = \sum_{j=1}^{n_i} y_{ij}/n_i = \bar{y}_{i\cdot}$. The ML solution coincides, since maximizing the likelihood function is equivalent to minimizing the quantity Q at the exponent. We need also an estimate for the parameter σ^2, which is not of direct interest, but still is needed in further steps of the analysis. An appropriate estimator for σ^2 will be given in the next sections. Finally, we can define the residuals, as $e_{ij} = y_{ij} - \hat{\mu}_i$, $\forall i = 1, \ldots, r$ and $\forall j = 1, \ldots, n_i$. They represent the deviation of y_{ij} around the estimated factor level mean. Residuals can be used to examine the aptness of the model for a given data set.

2.4 SSTO, SSTR, SSE

A fundamental identity

We can partition the total variability of the observations in difference parts:

- this practice can help us in recognizing difference sources of variability and assigning to each of them its relative importance.
- this will be then useful to understand how much the fitted ANOVA model is able to improve the simplest possible model (which assumes an overall common mean for all the factor levels)

In practice:

- the total variability of the observations is measured in terms of the total deviation of the observations around the overall mean $(y_{ij} - \bar{y}_{\cdot\cdot})$.
- Once we fit the ANOVA model, we can decompose this total deviation in
  - the deviation of the observations from their specific factor levels estimates $(y_{ij} - \bar{y}_{i\cdot})$ and
  - the remaining variability of the factor levels estimates from the overall mean $(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$.

Thus we obtain:

$$y_{ij} - \bar{y}_{\cdot\cdot} = (y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) .$$

If now we take the square of both sides of the equation and then sum, we obtain:

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left[(y_{ij}-\bar{y}_{i\cdot})+(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})\right]^2$$

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2 + \sum_{i=1}^{r} n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i} 2\left[(y_{ij}-\bar{y}_{i\cdot})(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})\right]$$

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2 + \sum_{i=1}^{r} n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 + \underbrace{2\sum_{i=1}^{r}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})}_{0}$$

$$\underbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\cdot\cdot})^2}_{\text{Total sum of squares } (SSTO)} = \underbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2}_{\text{Error sum of squares } (SSE)} + \underbrace{\sum_{i=1}^{r} n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}_{\text{Treatment sum of squares } (SSTR)}$$

Variance components' degrees of freedom

We can also easily obtain the degrees of freedom (*df*, number of independent/free observations for the estimation) associated to this variance decomposition

- $SSTO$ has $n_T - 1$ *df*, because 1 *df* is lost because of the constrain $\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\cdot\cdot}) = 0$.
- $SSTR$ has $r - 1$ *df*, because 1 *df* is lost because of the constrain $\sum_{i=1}^{r} n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}) = 0$.
- SSE has $n_T - r$ *df*, because r *df* are lost because each component of SSE, for the i^{th} factor, has the constrain $\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot}) = 0$, and thus 1 *df* is lost for each component.

Mean squares

The mean squares are obtained dividing each sum of squares by its associated *df*, thus

$$\underbrace{MSTR = \frac{SSTR}{r-1}}_{\text{Treatment mean square}} \qquad \text{and} \qquad \underbrace{MSE = \frac{SSE}{n_T - r}}_{\text{Error mean square}}$$

These mean squares represent the average squared deviations, hence they are basically variance estimates (recall that $V[Z] = E[(Z - E[Z])^2]$).

The ANOVA table

To summarize:

Source of Variation	**SS**	**df**	**MS**	E[**MS**]
Between treatments	$\overbrace{\sum_{i=1}^{r} n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}^{SSTR}$	$r-1$	$\overbrace{\frac{SSTR}{r-1}}^{MSTR}$	$\sigma^2 + \frac{\sum_{i=1}^{r} n_i(\mu_i-\mu_\cdot)^2}{r-1}$
Error (within treatments)	$\overbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2}^{SSE}$	$n_T - r$	$\overbrace{\frac{SSE}{n_T - r}}^{MSE}$	σ^2
Total	$\overbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\cdot\cdot})^2}^{SSTO}$	$n_T - 1$		

Attention to the $E[MS]$: motivation to build the test!

R example: $SSTO$, $SSTR$, SSE

```
> ### Estimating $\mu_i$, $\sigma^2$, calculating SSE,SSTR and SSTO
> # means
> mu_i = c(mean(y[group==1]),mean(y[group==2]),mean(y[group==3]))
> mu_i
[1] 6.877778 8.133333 9.200000
>
> # SSE,SSTR and SSTO
> sum_i1_sstr = length(y[group==1])*((mu_i[1]-mean(y))^2)
> sum_i2_sstr = length(y[group==2])*((mu_i[2]-mean(y))^2)
> sum_i3_sstr = length(y[group==3])*((mu_i[3]-mean(y))^2)
> sum_i_sstr = c(sum_i1_sstr,sum_i2_sstr,sum_i3_sstr)
> sum_i_sstr
[1] 10.3827160  0.3952263  9.3472428
> sstr = sum(sum_i_sstr )
> sstr
[1] 20.12519
> sum_i1_sse = sum((y[group==1]-mu_i[1])^2)
> sum_i2_sse = sum((y[group==2]-mu_i[2])^2)
> sum_i3_sse = sum((y[group==3]-mu_i[3])^2)
> sum_i_sse = c(sum_i1_sse,sum_i2_sse,sum_i3_sse)
> sum_i_sse
[1] 5.295556 6.306667 3.760000
> sse = sum(sum_i_sse)
> sse
[1] 15.36222
>
```

```
> ssto = sum((y-mean(y))^2)
> ssto
[1] 35.48741
>
> # $\sigma^2$
> sigma2 = sse/(length(y)-3)   # we'll see it later
> sigma2
[1] 0.6400926
```

2.5 F statistic

Expected mean squares

The expected mean squares are

$$E[MSE] = \sigma^2 \qquad \text{and} \qquad E[MSTR] = \sigma^2 + \frac{\sum_{i=1}^{r} n_i \left(\mu_i - \mu_\cdot\right)^2}{r-1}$$

Notice that:

- $MSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2 / \left(n_T - r\right) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} e_{ij}^2 / \left(n_T - r\right)$ is an unbiased estimator of $\sigma^2$, which is, in the ANOVA model, the variance of the errors ϵ_{ij}.
- $E[MSTR] = \sigma^2$ when all the means $\mu_i$ are equal, which means under the H_0. Instead $E[MSTR] > \sigma^2$ when at least one of the means $\mu_i$ is different, which means when H_0 does not hold.

This means that by comparing $E[MSTR]$ and $E[MSE]$ we can have information about the truth of H_0, since:

- when $E[MSTR] = E[MSE]$, then H_0 is true,
- when $E[MSTR] > E[MSE]$.

Hence, we can test H_0 versus H_1 by comparing these two quantities. Obviously, $E[MSTR]$ and $E[MSE]$ are unknown, and we will need to work with their observed versions $MSTR$ and MSE. Thus the ANOVA problem is solved, we just need to:

(1) find a smart way (a test statistic) to compare $MSTR$ and MSE,

(2) recover its distribution under H_0,

(3) perform a standard statistical testing procedure.

F statistic

Let us recall the tested hypotheses:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_r \\ H_1: & \text{at least one } \mu_i \text{ is different} \end{cases} .$$

We just saw that we can get information about the truth of H_0 by comparing $MSTR$ and MSE. In practice, the test statistic to be used is:

$$F = \frac{MSTR}{MSE}$$

Large values of F support H_1, since they indicate that $MSTR > MSE$. Values of F close to 1 support H_0, since they indicate $MSTR \approx MSE$. As for any statistical test, we need to set a decision rule: how large does the observed value for F, F^{obs}, need to be to reject H_0? The general rule is: we recover the distribution of F under H_0, and we check if F^{obs} belongs to the rejection region R for the test. In our case:

- when H_0 holds, $F \sim F_{(r-1,n_T-r)}$,
- we reject H_0 when F^{obs} takes large values, which means that our rejection region R is on the right tail of the distribution. For a fixed level of significance α, $R = \{F : F^{obs} > F_{(1-\alpha;r-1,n_T-r)}\}$.

R example: applying the F test

We can compute the test statistic by hand, or use R functions:

```
> ### Applying the F test
>
> F_obs = (sstr/(3-1))/(sse/(length(y)-3))
> F_obs
[1] 15.72053
> F_crit = qf(0.95,3-1,length(y)-3)
> F_crit
[1] 3.402826
> p_value = 1-pf(F_obs,3-1,length(y)-3)
> p_value
[1] 4.330692e-05
>
>
> # Using R functions
> mod = lm(y~as.factor(group))
> anova(mod)
Analysis of Variance Table

Response: y
                 Df Sum Sq Mean Sq F value    Pr(>F)    
as.factor(group)  2 20.125 10.0626   15.72 4.331e-05 ***
Residuals        24 15.362  0.6401                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Randomization F test

Randomization (or permutation) tests are distribution free procedures to test statistical hypothesis. A randomization test can be built starting from any parametric test, provided that a randomization scheme is used in the study. Any randomization test works according to the following steps (exactly as for the t-test):

(1) choose a suitable test statistic T (again, T is a random variable),

(2) compute T on the observed dataset, getting the value t^{obs},

(3) compare t^{obs} with the distribution of T under H_0,

(4) reject H_0 if and only if $t^{obs} \in R$, where R is the rejection region for the test (at some fixed significance level).

The difference between a parametric and a randomization test is in step (3), when we compare t^{obs} with the distribution of T under H_0. In this case no assumptions are made on the distribution of the data, hence, in general, it is not possible to recover any parametric distribution for the test statistic T. The problem is solved estimating the exact randomization distribution of T. This is obtained by exploiting the randomization scheme used to assign subjects to factor levels:

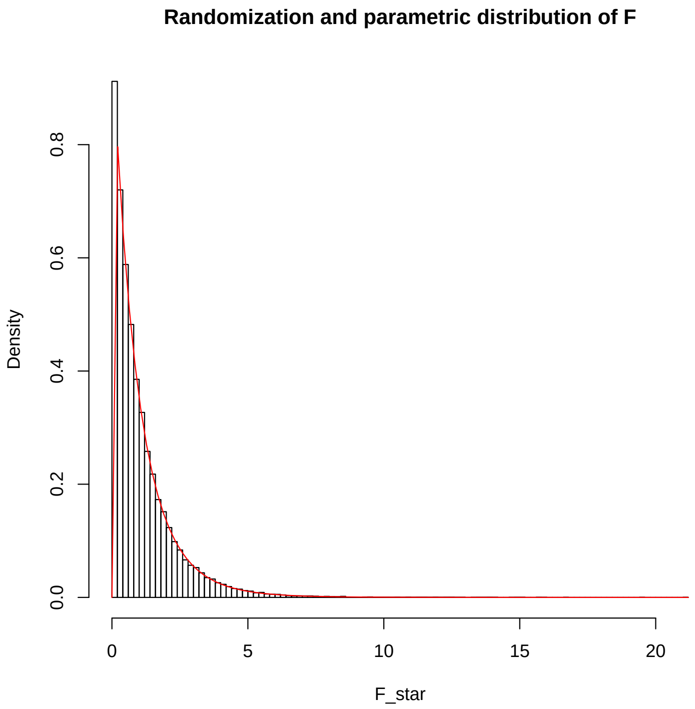
- under the null hypothesis the response Y have the same distribution $\mathcal{F}$ for all the factor levels,
- thus the observed data would have been the same if another randomization would have been applied.
- this means that we can sample from the null distribution of T just by re-assigning the experimental units to the factor levels and compute the test statistic on the re-randomized dataset.
- by repeating the previous step a large number B of times, we end up with a sample of values for T (descrete randomization distribution, $t^* = \{t^{*,b}\}_{b=1,\ldots,B}$), which we can use to estimate its exact distribution under H_0,
- we can then define the rejection region R by referring to the tails of the randomization distribution of T;
- and the p-value can be computed, for instance if large values of T support H_1, as $\lambda = (\#t^* > t^{obs})/B$.

R example: randomization F test

Let us perform the randomization F test

```
> ### Randomization F test
> # preparing the objects
> B = 100000
> F_star = array(0,dim=c(B))
> # re-randomizing the data and computing F of the new datasets
> for(i in 1:B){
```

```
+ group_new = sample(group,size=length(group),replace=F)
+ mod_star = lm(y~as.factor(group_new))
+ F_star[i] = anova(mod_star)$F[1]
+ }
>
> hist(F_star,breaks=100,prob=T,main="Randomization and parametric distribution of F")
> curve(df(x,3-1,length(y)-3),add=T,col="red")
>
> rand_p_value = sum(F_star>F_obs)/B
> rand_p_value
[1] 6e-05
> param_p_value = anova(mod)$Pr[1]
> param_p_value
[1] 4.330692e-05
```



ANALYSIS OF VARIANCE

Master of Statistics

Lesson 2

Dr. Francesca Solmi

Contents

1 Analysis of Factor Level Effects

Factor levels effects

Kutner et al. (2005), Chapter 17.

We are working with the ANOVA model

$$y_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, \dots, r$ and $j = 1, \dots, n_i$. In the last lesson we saw how to perform the F test to test whether the means μ_i differ. This test tells us only if there is at least one of the means μ_i that differs from the others, BUT it does not tell us which are these means, and how they differ from the others. Today we will see how a further analysis of the factor levels means can be performed, once the F test refuses H_0.

1.1 Graphical means

Graphical means

Let us recall the productivity improvement example from the previous lesson:

```
> ### Applying the F test
>
> F_obs = (sstr/(3-1))/(sse/(length(y)-3))
> F_obs
[1] 15.72053
> F_crit = qf(0.95,3-1,length(y)-3)
> F_crit
[1] 3.402826
> p_value = 1-pf(F_obs,3-1,length(y)-3)
> p_value
[1] 4.330692e-05
>
>
> # Using R functions
> mod = lm(y˜as.factor(group))
> anova(mod)
Analysis of Variance Table

Response: y
                 Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(group)  2 20.125 10.0626   15.72 4.331e-05 ***
Residuals        24 15.362  0.6401
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

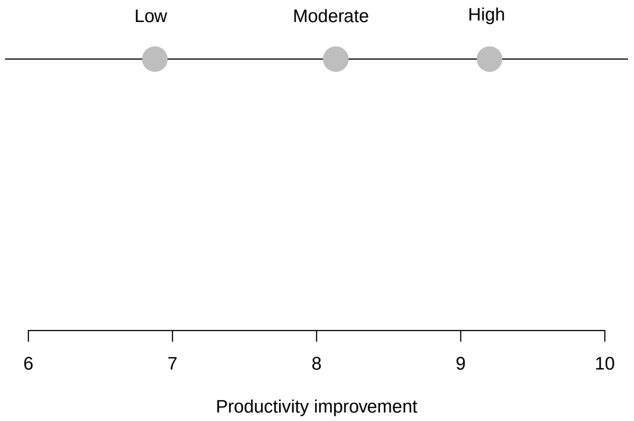
The F test refuses H_0, thus a further analysis on the factor levels means is needed. There are three simple graphical methods to display the factor levels means:

- *line plot*: it simply shows the position of the sample means $\bar{y}_{i\cdot}$ on a straight line,
- *bar graph*: it uses vertical bars to display the sample means $\bar{y}_{i\cdot}$,
- *main effects plot*: it reports the sample means $\bar{y}_{i\cdot}$ in a scatter plot.

Line plot

We just have to display the sample means $\bar{y}_{i\cdot}$ on a straight line.

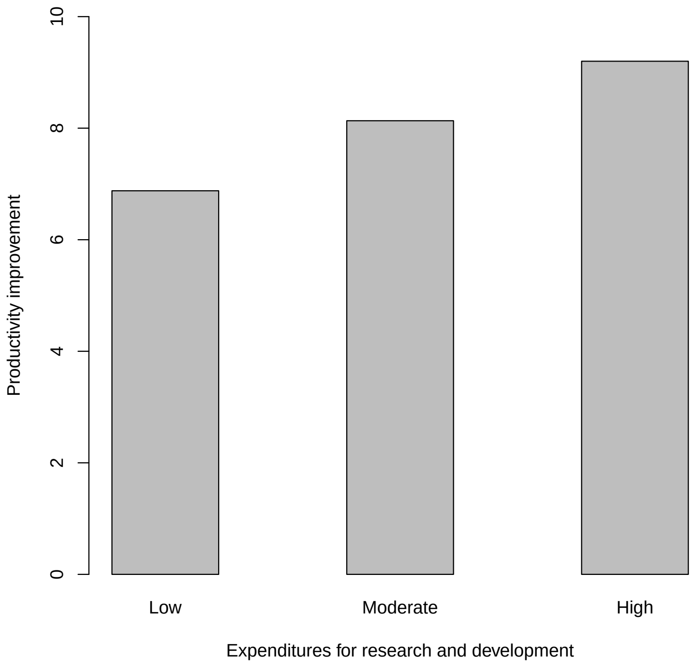
```
> # line plot
>
> plot(mu_i,c(1,1,1),
+      axes=F,pch=19,cex=3,
+      ylim=c(0,2.100),xlim=c(6,10),
+      xlab="Productivity improvement",
+      ylab=" ",col="grey")
> axis(1)
> abline(h=1)
> points(mu_i,c(1,1,1),pch=19,cex=3,col="grey")
> text(6.85,1.3,"Low")
> text(8.1,1.3,"Moderate")
> text(9.18,1.3,"High")
```



Bar graph

It is a graphic in two dimensions. We use vertical bars to display the factor level means.

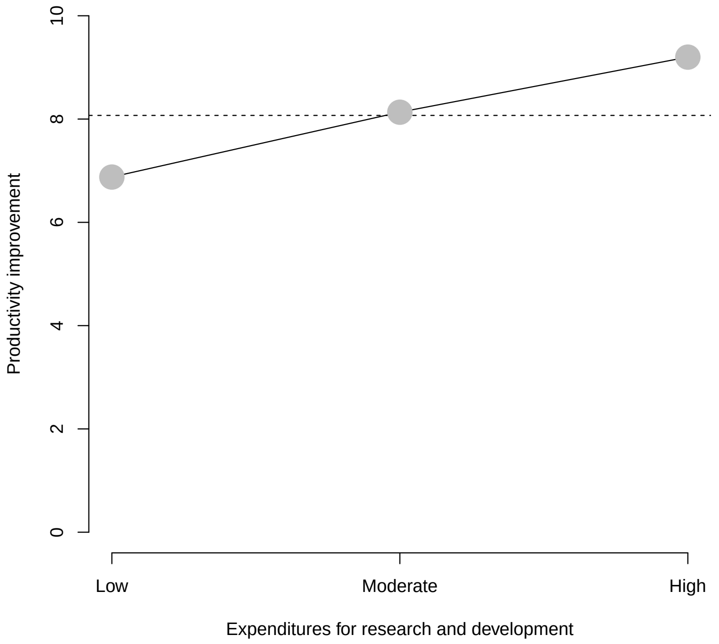
```
> barplot(height=mu_i,width=1,space=1.2,
+      ylim=c(0,10),
+      names.arg=c("Low","Moderate","High"),
+      xlab="Expenditures for research and development",
+      ylab="Productivity improvement",col="grey")
```



Main effects plot

It is a graphic in two dimensions. We provide a scatter plot of the factor level means, and then connect the circles through straight lines (to underline possible differences).

```
> # main effects plot
>
> plot(c(1,2,3),mu_i,
+      axes=F,
+      ylim=c(0,10),
+      xlab="Expenditures for research and development",
+      ylab="Productivity improvement",type="l")
> abline(h=mean(mu_i),lty="dashed")
> points(c(1,2,3),mu_i,pch=19,cex=3,col="grey")
> axis(1,at=c(1,2,3),labels=c("Low","Moderate","High"))
> axis(2)
```



1.2 Estimation and testing

Estimation and testing procedures

Several specific quantities may be of interest after the F test refuses H_0:

- a single factor level mean μ_i,
- a difference between two factor level means,
- a contrast among factor level means,
- a linear combination of factor level means.

Difference inferential procedures must be implemented according to which is our objective.

Single factor level mean

We know already that an unbiased point *estimate* for μ_i is $\hat{\mu}_i = \bar{y}_{i\cdot}$. The related *estimator* $\bar{Y}_{i\cdot}$ is characterized by:

- having mean $E\left[\bar{Y}_{i\cdot}\right] = \mu_i$,
- having variance $V\left[\bar{Y}_{i\cdot}\right] = \frac{\sigma^2}{n_i}$,
- having normal distribution.

We can use this information to build an interval estimate for μ_i: this can be done by exploiting the distribution of a particular statistic built on $\bar{Y}_{i\cdot}$ (as for building a test). We can use the (already known) fact that an unbiased *estimator* for σ^2 is MSE. Then an estimator $S^2_{\bar{Y}_{i\cdot}}$ for the variance of $\bar{Y}_{i\cdot}$ is given by replacing σ^2 with MSE:

$$S^2_{\bar{Y}_{i\cdot}} = \frac{MSE}{n_i}.$$

Then the statistic $(\bar{Y}_{i\cdot} - \mu_i)/S_{\bar{Y}_{i\cdot}}$ has distribution $t_{n_T - r}$. Thus we can provide a *confidence interval* (CI)

for μ_i as

$$\bar{y}_{i\cdot} \pm t_{1-\alpha/2;n_T-r} s_{\bar{Y}_{i\cdot}}$$

Such CI can also be used to test the hypotheses

$$\begin{cases} H_0: & \mu_i = c \\ H_1: & \mu_i \neq c \end{cases},$$

where c is some constant of interest. We conclude that H_0 cannot be rejected when c is included in the CI. Equivalently we can use the test statistic $(\bar{Y}_{i\cdot} - c)/S_{\bar{Y}_{i\cdot}}$, which follows a t_{n_T-r} distribution under H_0. Let us go back to our example: we want to test $H_0 : \mu_1 = 8$:

```
## Single factor level mean
> alpha = 0.05
> n_T = length(data$y)
> r = length(mu_i)
> n_1 = sum(data$group==1)
> c = 8
>
> # confidence interval
>
> ci_mu1_single =  c(mu_i[1]+qt(alpha/2,n_T-r)*sqrt(sigma2/n_1),
+ mu_i[1]+qt(1-alpha/2,n_T-r)*sqrt(sigma2/n_1))
> ci_mu1_single
[1] 6.327365 7.428191
>
> # two-sided test (H_1: mu_1 \neq 8)
>
> twosided_pvalue_mu1_single = min(1-pt((mu_i[1]-c)/sqrt(sigma2/n_1),n_T-r),
+                         pt((mu_i[1]-c)/sqrt(sigma2/n_1),n_T-r))*2
> twosided_pvalue_mu1_single
[1] 0.0003111013
>
> # one-sided test (H_1: mu_1 < 8)
>
> onesided_pvalue_mu1_single = pt((mu_i[1]-c)/sqrt(sigma2/n_1),n_T-r)
> onesided_pvalue_mu1_single
[1] 0.0001555506
```

Difference between two factor level means

We can obtain an unbiased point *estimate* for $D = \mu_i - \mu_{i'}$ as $\hat{d} = \bar{y}_{i\cdot} - \bar{y}_{i'\cdot}$. This is called *pairwise comparison*. The related *estimator* $\hat{D}$ is characterized by:

- having mean $E\left[\hat{D}\right] = \mu_i - \mu_{i'}$,
- having variance $V\left[\hat{D}\right] = V\left[\bar{Y}_{i\cdot}\right] + V\left[\bar{Y}_{i'\cdot}\right] = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)$,
- having normal distribution, since linear combination of independent, normally distributed, random variables.

We can use this information to build an interval estimate for D: this can be done, again by exploiting the distribution of a particular statistic built on $\bar{D}$ (as for building a test). We can use again the fact that an unbiased *estimator* for σ^2 is MSE. Then an estimator $S^2_{\hat{D}}$ for the variance of $\hat{D}$ is given by replacing σ^2 with MSE:

$$S^2_{\bar{D}} = MSE\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right).$$

Then the statistic $(\hat{D} - D)/S_{\hat{D}}$ has distribution $t_{n_T - r}$ (proof analogous to the previous one). Thus we can provide a CI for D as

$$\hat{d} \pm t_{1-\alpha/2;n_T-r} s_{\hat{D}}$$

and it can also be used to test the hypotheses

$$\begin{cases} H_0: & D = 0 \\ H_1: & D \neq 0 \end{cases},$$

We conclude that H_0 cannot be rejected when 0 is included in the CI. Equivalently we can use the test statistic $\hat{D}/S_{\hat{D}}$, which follows a t_{n_T-r} distribution under H_0. Let us go back to our example: we want to test $H_0: \mu_2 - \mu_1 = 0$:

```
> ## Pairwise comparisons
>
> n_2 = sum(data$group==2)
> n_3 = sum(data$group==3)
>
> # confidence interval for D = mu_2 - mu_1
>
> dmu1mu2 = mu_i[2]-mu_i[1]
> dmu1mu2
[1] 1.255556
> ci_Dmu1mu2 =  c(dmu1mu2+qt(alpha/2,n_T-r)*sqrt(sigma2*(1/n_1+1/n_2)),
+ dmu1mu2+qt(1-alpha/2,n_T-r)*sqrt(sigma2*(1/n_1+1/n_2)))
> ci_Dmu1mu2
[1] 0.5274279 1.9836832
>
> # two-sided test (H_1: mu_2 - mu_1 \neq 0)
>
> twosided_pvalue_Dmu1mu2 = min(1-pt((dmu1mu2)/sqrt(sigma2*(1/n_1+1/n_2)),n_T-r),
+                                 pt((dmu1mu2)/sqrt(sigma2*(1/n_1+1/n_2)),n_T-r))*2
> twosided_pvalue_Dmu1mu2
[1] 0.001592147
>
> # one-sided test (H_1: mu_2 - mu_1 > 0)
>
> onesided_pvalue_mu1_single = 1-pt((dmu1mu2)/sqrt(sigma2*(1/n_1+1/n_2)),n_T-r)
> onesided_pvalue_mu1_single
[1] 0.0007960733
```

Contrasts of factor level means

A *contrast* is a comparison involving two or more factor level means (the pairwise comparison case is then included in this). A contrast is in general defined as a linear combination of the factor level means μ_i, where the coefficients c_i sum to 0:

$$L = \sum_{i=1}^{r} c_i \mu_i, \quad \text{where } \sum_{i=1}^{r} c_i = 0.$$

Different contrasts can be generated (in our example):

- $L = \mu_1 - \mu_2$: comparison between low and moderate expenditures,
- $L = \mu_3 - \frac{\mu_1+\mu_2}{2}$: comparison between high expenditures vs low and moderate (not high),
- $L = \mu_3 - \frac{\mu_1+\mu_2+\mu_3}{3}$: comparison between high expenditures and the average (effect of high $\tau_3 = \mu_3 - \mu_.$).

We can obtain an unbiased point *estimate* for L as $\hat{l} = \sum_{i=1}^{r} c_i \bar{y}_{i\cdot}$. The related *estimator* $\hat{L}$ is characterized by:

- having mean $E\left[\hat{L}\right] = \sum_{i=1}^{r} c_i \mu_i$,
- having variance $V\left[\hat{L}\right] = \sum_{i=1}^{r} c_i^2 V\left[\bar{Y}_{i\cdot}\right] = \sum_{i=1}^{r} c_i^2 \left(\frac{\sigma^2}{n_i}\right) = \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}$,
- having normal distribution, since linear combination of independent, normally distributed, random variables.

Again, we can use this information to build an interval estimate for L: this can be done by exploiting the distribution of a particular statistic built on $\hat{L}$ (as for building a test). We can use again the fact that an unbiased *estimator* for σ^2 is MSE. Then an estimator $S^2_{\hat{L}}$ for the variance of $\hat{L}$ is given by replacing σ^2 with MSE:

$$S^2_{\hat{L}} = MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}.$$

Then the statistic $(\hat{L} - L)/S_{\hat{L}}$ has distribution $t_{n_T - r}$ (proof analogous to the previous ones). Thus we can provide a CI for L as

$$\hat{l} \pm t_{1-\alpha/2; n_T - r} s_{\hat{L}}$$

and it can also be used to test the hypotheses

$$\begin{cases} H_0 : & L = 0 \\ H_1 : & L \neq 0 \end{cases},$$

We conclude that H_0 cannot be rejected when 0 is included in the CI. Equivalently we can use the test statistic $\hat{L}/S_{\hat{L}}$, which follows a $t_{n_T - r}$ distribution under H_0. Let us go back to our example: we want to test $H_0 : \mu_3 = \mu$:

```
> ## Contrasts
>
> c_1 = -1/3
> c_2 = -1/3
> c_3 = 2/3
> c_i = c(c_1,c_2,c_3)
> n_i = c(n_1,n_2,n_3)
>
> # confidence interval for L = mu_3-(mu_1+mu_2+mu_3)/3 (delta_3 effect)
>
> ldelta3 = mu_i[3]-(mu_i[1]+mu_i[2]+mu_i[3])/3
> ldelta3
[1] 1.12963
> ci_ldelta3 =  c(ldelta3+qt(alpha/2,n_T-r)*sqrt(sigma2*sum((c_i^2)/n_i)),
+                 ldelta3+qt(1-alpha/2,n_T-r)*sqrt(sigma2*sum((c_i^2)/n_i)))
> ci_ldelta3
[1] 0.6188682 1.6403911
>
> # two-sided test (H_1: mu_3 \neq mu)
>
> twosided_pvalue_ldelta3 = min(1-pt((ldelta3)/sqrt(sigma2*sum((c_i^2)/n_i)),n_T-r),
+                               pt((ldelta3)/sqrt(sigma2*sum((c_i^2)/n_i)),n_T-r))*2
> twosided_pvalue_ldelta3
[1] 0.0001256483
>
```

```
> # one-sided test (H_1: mu_3 > mu)
>
> onesided_pvalue_ldelta3 = 1-pt((ldelta3)/sqrt(sigma2*sum((c_i^2)/n_i)),n_T-r)
> onesided_pvalue_ldelta3
[1] 6.282413e-05
```

Linear combinations of factor level means

We may also be interested in linear combinations of the factor level means that are not contrasts (the c_i do not sum to 0). We can perform estimation and testing following the same theory for contrasts. In our example, this could be the case if we are interested in estimating the overall mean productivity improvement of the same firms for next year, when some of them will change their expenditures for research. Assume next year only 5 of them will have low level of expenditures, only 8 will have moderate level and the remaining 14 will have high level. Then we would be interested in:

$$L = \frac{5}{27}\mu_1 + \frac{8}{27}\mu_2 + \frac{14}{27}\mu_3,$$

which is a linear combination of the μ_i where the coefficients c_i sum to 1 and not to 0.

1.3 Multiple comparisons procedures

Simultaneous inference procedures

The estimation and testing procedures we just saw have two important limitations:

- **(Global type I error rate)** the confidence coefficient $1-\alpha$ applies only to the single estimates, and not to a series of estimates. Hence, if more that one CI is calculated or test is performed, the type I error rate is α for the single inference, but not for the *global* procedure,
- **(data snooping)** the confidence coefficient $1-\alpha$ and the significance level α are appropriate only if the estimate or test was not suggested by the data.

For this reason more suitable procedure should be used when the inference on the factor level means is *multiple*. There is, in this case, the need of a simultaneous inference procedure.

Global type I error rate

Any statistical test is built taking care of two inferential errors:

- the concept of type I error is related to the probability that a test rejects the null hypothesis when this is true. When a test is built, this probability is kept fixed at a small value α,
- then the several existing tests are ranked in terms of their II type error β, or better in terms of their power $1-\beta$, which is defined as the probability that the test rejects a false null hypothesis. The higher the power the better the test.

This basic rule is used for the construction of a single test, but the same idea applies when we want to perform several tests at the same time. We can see the several separated tests as composing a global procedure, which is testing the hypotheses

$$\begin{cases} H_{0\,glob}: & \bigcap_{m=1}^{M} H_{0\,m} \\ H_{1\,glob}: & \bigcup_{m=1}^{M} H_{1\,m} \end{cases},$$

where $H_{0\,m}$ and $H_{1\,m}$, for $m = 1, \ldots, M$ are the null and alternative hypotheses tested by the partial tests. In this setting, we can then also define the inferential errors for the global procedure. In particular, let us focus on the confidence level $1-\alpha$, defined as the probability of not rejecting the null hypothesis when this is true. For the global test this becomes

$$\begin{aligned} Pr\left\{\text{not reject } H_{0\,glob}\right\} =& Pr\left\{\text{not reject } \bigcap_{m=1}^{M} H_{0\,m}\right\} \\ =^{*}& \prod_{m=1}^{M} Pr\left\{\text{not reject } \bigcap_{m=1}^{M} H_{0\,m}\right\} \\ =^{**}& (1-\alpha)^{M} < 1-\alpha, \end{aligned}$$

where the first equality (*) holds in case of independent partial tests, and the last equality (**) holds when we perform each of the partial tests at a $1-\alpha$ confidence level.

This means that the confidence levels of the partial tests need to be somehow increased in order to get a global procedure which works at the actual $1-\alpha$ confidence level. In other words, the significance levels of the partial tests need to be reduced before comparing the observed p-values. In other words (always the same concept) the p-values of the partial tests need to corrected before being compared to the α value. This is exactly the objective of the so-called *Multiple comparisons and/or tests procedures.*

Data snooping

In an experiment or an observational study, data snooping is a term often used to indicate the process of studying effects that are suggested by the data. For instance this is the case when, in an ANOVA setting, the factor level means associated to the lowest and the highest estimates are compared. The actual significance level for such a test can be much higher than the nominal one. The reason for this is related to what we have just seen about the confidence level of a global test: in this case, indeed, a family of multiple tests is conducted implicitly by the analyst (he would not know which levels means to compare, before the experiment starts). To solve this problem, we can apply a *multiple comparisons or tests procedure*, so to take into account, explicitly this time, all the needed partial tests.

Tukey procedure

This procedure can be applied when the family of interest is the set of *all pairwise comparisons* of factor levels means. Hence the global hypotheses are

$$\begin{cases} H_{0\,glob}: & \bigcap_{m=1}^{M} H_{0\,m} \\ H_{1\,glob}: & \bigcup_{m=1}^{M} H_{1\,m} \end{cases} = \begin{cases} H_{0\,glob}: & \bigcap_{i,i'=1}^{r} \mu_i - \mu_i' = 0 \\ H_{1\,glob}: & \bigcup_{i,i'=1}^{r} \mu_i - \mu_i' \neq 0 \end{cases},\ i \neq i'.$$

When all n_i are equal, then the procedure has an exact confidence level of $1-\alpha$. Instead, when some of the n_i are different, the procedure is characterized by having an actual confidence level higher than $1-\alpha$ (conservative procedure). The Tukey procedure is based on the *studentized range distribution*. Suppose r independent observations $T_1, \ldots, T_r$ from a $N(\mu, \sigma^2)$ distribution. We can define the range w for this set of observations as

$$w = max_{i=1,\ldots,r}\left\{T_i\right\} - min_{i=1,\ldots,r}\left\{T_i\right\}.$$

Suppose that an estimate for σ^2, s^2, is available, based on v degrees of freedom and it is independent of the T_i. Then the ratio w/s is called *studentized range*:

$$q\left(r, v\right) = \frac{w}{s}.$$

The studentized range has a tabulated distribution, which depends on r and v. This distribution can be used to perform simultaneous estimation of all pairwise comparisons. We can build a CI for any pairwise comparison as (after some calculations, which are approximate for unbalanced designs):

$$\hat{d} \pm \frac{1}{\sqrt{2}} q_{(1-\alpha;r,n_T-r)} s_{\hat{D}},$$

where $q_{(1-\alpha;r,n_T-r)}$ is the $(1-\alpha)\%$ quantile of a studentized range distribution. Here the the $1-\alpha$ level is related to the family of all pairwise comparisons, and not to the single ones. We can also perform a test for the null hypothesis $H_0 : D = 0$, and reject it if (for a two-sided alternative)

$$\left| \frac{\sqrt{2}\hat{d}}{s_{\hat{D}}} \right| > q_{(1-\alpha;r,n_T-r)},$$

It is also possible to construct CIs around the μ_i that are also controlling the global confidence level at the nominal level. Such CIs are given by (we need to correct the width of the interval, by dividing it by 2)

$$\bar{y}_{i\cdot} \pm \frac{q_{(1-\alpha;r,n_T-r)}}{2\sqrt{2}} s_{\hat{D}},$$

It is also possible to test the all the pairwise comparisons at the same time by graphical means. We can do this by displaying, in one of the graphics we saw before, the estimated $\bar{y}_{i\cdot}$ and $\bar{y}_{i'\cdot}$ and the related CIs, as just described. We can test which couples of factor levels means are significantly different by checking which couples of CIs are not overlapping.

Scheffé procedure

This procedure can be applied when the family of interest is the set of *all possible contrasts* of factor levels means. Hence the global hypotheses are

$$\begin{cases} H_{0\,glob}: & \bigcap_{m=1}^{M} H_{0\,m} \\ H_{1\,glob}: & \bigcup_{m=1}^{M} H_{1\,m} \end{cases} = \begin{cases} H_{0\,glob}: & \bigcap_{m=1}^{M_L} L_m = 0 \\ H_{1\,glob}: & \bigcup_{m=1}^{M_L} L_m \neq 0 \end{cases}.$$

Thus infinitely many statements belong to this family. The procedure has an exact confidence level of $1-\alpha$ whether the factor levels sample sizes are equal or not. The Sheffé procedure is based on the previous results obtained on the distribution of $\hat{L}$. After some calculations, we can build a CI for any pairwise comparison as:

$$\hat{l} \pm \sqrt{(r-1)\, F_{(1-\alpha;r-1,n_T-r)}}\, s_{\hat{L}},$$

where $F_{(1-\alpha;r-1,n_T-r)}$ is the $(1-\alpha)\%$ quantile of a F_{r-1,n_T-r} distribution. Here the $1-\alpha$ level is related to the family of all pairwise comparisons, and not to the single ones. We can also perform a test for the null hypothesis $H_0 : L = 0$, and reject it if (for a two-sided alternative)

$$\left| \frac{\hat{l}^2}{(r-1)\, s_{\hat{L}}^2} \right| > F_{(1-\alpha;r-1,n_T-r)},$$

The procedure takes into account all the possible contrasts, so it is conservative if a small number of them needs to be performed. In other words, it will show in this case a low power compared to other less conservative procedures. On the other end, it allows to choose the comparisons to be made after the experiments is conducted. Thus it can be used for a wide variety of data snooping. Compared to the Tukey method, the latter is recommended when only the family of all pairwise comparisons is of interest (narrower CIs are produced). The Sheffé procedure has the property that if the F test on the global difference among the factor level means is significant, then at least one of the corrected tests on all the possible contrasts will be also significant. And no significance will be found by the procedure if the global F test is not significant.

Bonferroni procedure

This procedure can be applied when the family of interest is a set of *pairwise comparisons*, *contrasts* and/or *linear combinations* of factor levels means, which is specified by the analyst in advance of the data analysis. Hence the global hypotheses are

$$\begin{cases} H_{0\,glob}: & \bigcap_{m=1}^{M} H_{0\,m} \\ H_{1\,glob}: & \bigcup_{m=1}^{M} H_{1\,m} \end{cases} = \begin{cases} H_{0\,glob}: & \bigcap_{m=1}^{M_e} L_m = 0 \\ H_{1\,glob}: & \bigcup_{m=1}^{M_e} L_m \neq 0 \end{cases}.$$

where M_e denotes the total number of partial hypotheses that we want to test in the given experiment. The procedure has an approximate confidence level of $1-\alpha$ whether the factor levels sample sizes are equal or not. The Bonferroni procedure is based on the previous results obtained on the distribution of $\hat{L}$. We can build a CI for any pairwise comparison as:

$$\hat{l} \pm t_{(1-\alpha/(2M_e);n_T-r)} s_{\hat{L}},$$

Notice that the only difference with the non corrected CIs for L is the confidence level in $t_{(1-\alpha/(2M_e);n_T-r)}$, which is increased from $1-\alpha/2$ to $1-\alpha/(2M_e)$. We can also perform a test for the null hypothesis $H_0: L = 0$, and reject it if (for a two-sided alternative)

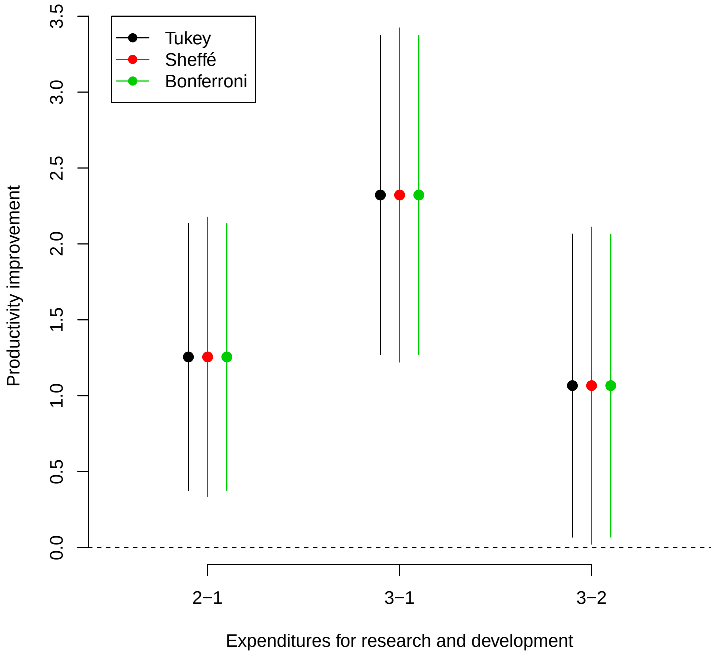
$$\left|\frac{\hat{l}}{s_{\hat{L}}}\right| > t_{(1-\alpha/(2M_e);n_T-r)},$$

The procedure can be quite conservative, when many linear combinations of factor levels means are tested, in the sense that it is possible that none of the partial tests rejects the partial null hypothesis, because too severe correction is applied. Compared to the Tukey method, the latter is recommended when the all the pairwise comparisons are of interest (narrower CIs are produced). When only some comparisons are of interest, then the two methods need to be compared case by case. Compared to the Sheffé method, the Bonferroni procedure is recommended when the number of contrasts of interest is about the same as the number of factor levels, or less. When not sure about which to choose, we can apply the three methods, and compare them. Let us go back to our example:

```
> ### Multiple comparisons procedures
> anova_table = aov(y~group1,data=data)
> summary(anova_table)
            Df Sum Sq Mean Sq F value   Pr(>F)
group1       2  20.12   10.06   15.72 4.33e-05 ***
Residuals   24  15.36    0.64
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> ## Tukey
> library(multcomp)
> data$group1 = as.factor(data$group)
> mod = lm(y~group1,data=data)
> tukey = glht(mod, linfct = mcp(group1 = "Tukey"))
> tukey_ci = confint(tukey)

> tukey_ci

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = y ~ group1, data = data)
```

```
Quantile = 2.4932
95% family-wise confidence level


Linear Hypotheses:
           Estimate lwr     upr
2 - 1 == 0 1.25556  0.37598 2.13513
3 - 1 == 0 2.32222  1.27093 3.37352
3 - 2 == 0 1.06667  0.06932 2.06401

> #plot(tukey)

> ## Sheffé
> all_pairwise = rbind("2 - 1" = c(-1, 1, 0),
+                  "3 - 1" = c(-1, 0, 1),
+                  "3 - 2" = c(0, -1, 1))
>
> sheffe_fun = function(y,mu_i,tested_contr,sse,n_T,r,alpha){
+ hat_l = NULL
+ s2_L = NULL
+ ci_L = array(0,dim=c(dim(tested_contr)[2],3))
+ pvalue_L0 = NULL
+ for(cont in 1:dim(tested_contr)[2]){
+ hat_l[cont] = (mu_i%*%tested_contrasts)[cont]
+ s2_L[cont] = (sse/(n_T-r))*sum(((tested_contrasts[,cont])^2)/n_i)
+ ci_L[cont,1:3] = c(hat_l[cont]-sqrt((r-1)*qf(1-alpha,r-1,n_T-r)*s2_L[cont]),
+                    hat_l[cont],
+                    hat_l[cont]+sqrt((r-1)*qf(1-alpha,r-1,n_T-r)*s2_L[cont]))
+ pvalue_L0[cont] = 1-pf(((hat_l[cont]^2)/((r-1)*s2_L[cont])),r-1,n_T-r)
+ }
+ list("p_value_sheffe"=pvalue_L0,"ci_sheffe"=ci_L)
+ }
>
> y = data$y
> tested_contrasts = t(all_pairwise)
> sheffe = sheffe_fun(y,mu_i,tested_contr,sse,n_T,r,alpha)

> sheffe
$p_value_sheffe
[1] 6.185819e-03 5.521957e-05 4.443457e-02

$ci_sheffe
           [,1]     [,2]     [,3]
[1,] 0.33520337 1.255556 2.175908
[2,] 1.22219096 2.322222 3.422253
[3,] 0.02308538 1.066667 2.110248
>
> ## Bonferroni
> bonferroni = glht(mod, linfct = mcp(group1 = all_pairwise), test = adjusted("bonferroni"))
> summary(bonferroni, test = adjusted("bonferroni"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts
```

```
Fit: lm(formula = y ~ group1, data = data)

Linear Hypotheses:
          Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0   1.2556     0.3528   3.559  0.00478 **
3 - 1 == 0   2.3222     0.4217   5.507 3.47e-05 ***
3 - 2 == 0   1.0667     0.4000   2.666  0.04051 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- bonferroni method)

> bonferroni_ci = confint(bonferroni)
> #plot(bonferroni)

> ## graphical comparison between the three methods
>
> plot(tukey_ci$confint[,1]~c(0.9,1.9,2.9),xlab="Expenditures for research and development",
+      ylab="Productivity improvement",axes=F,
+      ylim=c(min(tukey_ci$confint,sheffe$ci_sheffe,bonferroni_ci$confint),
+             max(tukey_ci$confint,sheffe$ci_sheffe,bonferroni_ci$confint)),
+      xlim=c(0.5,3.5))
> group_means = c(mean(y[group==1]),mean(y[group==2]),mean(y[group==3]))
> points(tukey_ci$confint[,1]~c(0.9,1.9,2.9),pch=19,cex=1.2)
> points(sheffe$ci_sheffe[,2]~c(1,2,3),pch=19,cex=1.2,col=2)
> points(bonferroni_ci$confint[,1]~c(1.1,2.1,3.1),pch=19,cex=1.2,col=3)
> axis(side=1, at = c(1,2,3), labels = c("Low","Moderate","High"))
> axis(side=2)
> abline(h=0,lty="dashed")
> segments(0.9,tukey_ci$confint[1,2],0.9,tukey_ci$confint[1,3])
> segments(1,sheffe$ci_sheffe[1,1],1,sheffe$ci_sheffe[1,3],col=2)
> segments(1.1,bonferroni_ci$confint[1,2],1.1,bonferroni_ci$confint[1,3],col=3)
> segments(1.9,tukey_ci$confint[2,2],1.9,tukey_ci$confint[2,3])
> segments(2,sheffe$ci_sheffe[2,1],2,sheffe$ci_sheffe[2,3],col=2)
> segments(2.1,bonferroni_ci$confint[2,2],2.1,bonferroni_ci$confint[2,3],col=3)
> segments(2.9,tukey_ci$confint[3,2],2.9,tukey_ci$confint[3,3])
> segments(3,sheffe$ci_sheffe[3,1],3,sheffe$ci_sheffe[3,3],col=2)
> segments(3.1,bonferroni_ci$confint[3,2],3.1,bonferroni_ci$confint[3,3],col=3)
> legend(0.5,3.5,pch=c(19,19,19),lty=c(1,1,1),col=c(1,2,3),
+        legend=c("Tukey","Sheffé","Bonferroni"))
```


Tukey
Sheffé
Bonferroni
Productivity improvement
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
2−1
3−1
3−2
Expenditures for research and development

ANALYSIS OF VARIANCE
Master of Statistics

Lesson 3

Dr. Francesca Solmi

Contents

1 Diagnostics and Remedial Measures

Choosing a good model

Kutner et al. (2005), Chapter 18.

So far, we saw how to do inference on an ANOVA model, but this is actually the final phase of the analysis. In fact, a suitable model should be chosen first, which is showing to be well fitting the data.

In practice, we should first check that the model is capturing the important features of the data. A way to do this is by analysis the *residuals* of the model. A specific ANOVA model can be not appropriate according to several aspects:

- nonconstancy of error variance σ^2,
- non independence of error terms,
- outliers
- omission of important explanatory variables
- nonnormality of error terms

All these aspects need to be studied, and, if needed, remedial measures must be adopted to increase the goodness of fit of the model.

1.1 Analysis of the residuals

Residuals definition

The analysis of the residuals can tell us many things about the goodness of fit of the model. Let us recall the ANOVA model

$$y_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, \dots, r$ and $j = 1, \dots, n_i$. where μ_i, $i = 1, \dots, r$, are the parameters and ϵ_{ij}, $i = 1, \dots, r$ and $j = 1, \dots, n_i$, are the random errors, assumed to be realizations of $N(0, \sigma^2)$. Notice that the distributional assumptions are actually on the errors of the model. The residuals of the models are defined as $e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot}$. Some transformations of the residuals are useful for diagnosing ANOVA model departures: *semistudentized residuals*, *studentized residuals*, *studentized deleted residuals*. *Semistudentized residuals* are defined as

$$e^*_{ij} = \frac{e_{ij}}{\sqrt{MSE}}.$$

Here the residuals e_{ij} are divided by $\sqrt{MSE}$, which is an approximation $\hat{\sigma}$ for their standard deviation. *Studentized residuals* are defined as

$$r_{ij} = \frac{e_{ij}}{s_{e_{ij}}}, \quad \text{where } s_{e_{ij}} = \sqrt{\frac{MSE\,(n_i - 1)}{n_i}}.$$

Here the residuals e_{ij} are divided by $s_{e_{ij}}$, which is an unbiased estimate of their standard deviation $\sigma_i = \sqrt{\sigma^2\,(1 - 1/n_i)}$, allowed in this case to be non constant for the different factor levels. *Studentized deleted residuals* are defined as

$$t_{ij} = e_{ij} \sqrt{\frac{n_T - r - 1}{SSE\left(1 - \frac{1}{n_i}\right) - e^2_{ij}}}.$$

These residuals are here represented as a function of e_{ij}. In fact, they are the residuals we would obtain by estimating the models each time on the whole dataset excluding y_{ij}, divided by an estimate of their standard deviation (for each factor level). Notice that, in balanced designs, the *semistudentized residuals* and the *studentized residuals* provide the same information, since they reduce to

$$\frac{e_{ij}}{\sqrt{MSE}} \quad \text{and} \quad \frac{e_{ij}}{\sqrt{\frac{MSE(n_T/r-1)}{n_T/r}}}.$$

Hence, they differ for the constant term $(1 - r/n_t)$, which is close to one if the sample size is not too low. Moreover notice that, in unbalanced designs, the errors associated to factor levels with a larger sample size n_i are standardized using a larger associate variance estimate. The meaning of the factor $(1 - 1/n_i)$ is very important: $1/n_i$ is the *leverage* of the error e_{ij}. It tells us how important is the observation y_{ij} in the model estimation, and hence how much importance we will have to give to the error e_{ij} when analyzing the residuals of the model. The idea is that, comparing different observations:

- if n_i is very large, then y_{ij} has a relatively low importance in the estimation of μ_i. In this case y_{ij} has a low *leverage*, and the associated studentized residual r_{ij} is rescaled to a lower number, in order to give it the right weight in the analysis.
- if n_i is very low, then y_{ij} has a relatively high importance in the estimation of μ_i. In this case y_{ij} has a high *leverage*, and the associated studentized residual r_{ij} is rescaled to a higher number.

Residual plots

We said before that the model validation can be done checking the good behavior of residuals. In particular, at a first stage, we can use residuals plots to check the following departures from the model:

- nonconstancy of error variance σ^2,
- non independence of error terms,
- outliers
- omission of important explanatory variables
- nonnormality of error terms

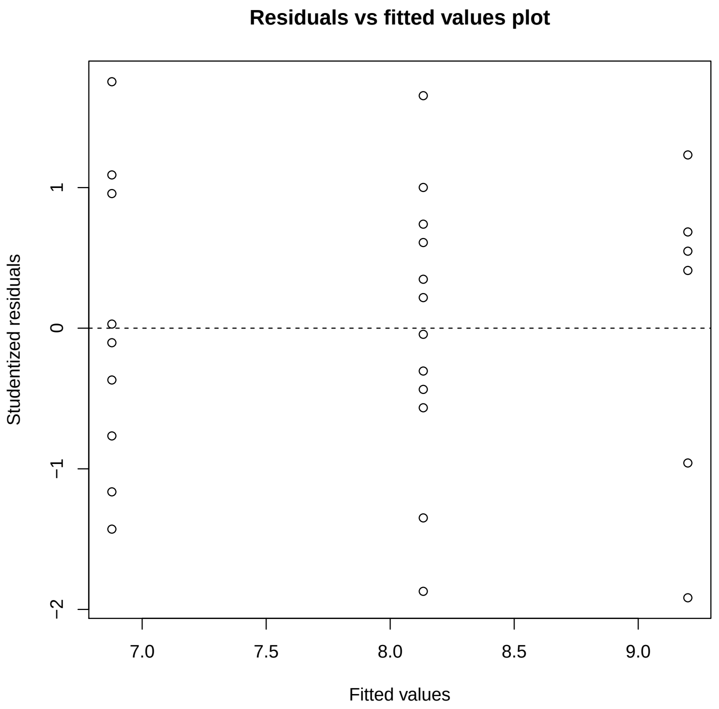
Several plots can be used to explore the residuals:

- plot against the fitted values $\hat{y}_{ij}$: we can detect possible nonconstancy of the error variance or the presence of outliers (use studentized residuals if the n_i differ a lot),
- time or other sequence plots: we can detect possible non independence of error terms (when the data are recorder over time),
- dot plots: we can detect possible nonconstancy of the error variance or the presence of outliers,
- normal probability plots: we can detect possible nonnormality of the error terms,
- any other plot which can be useful for detecting specific departures from the model.

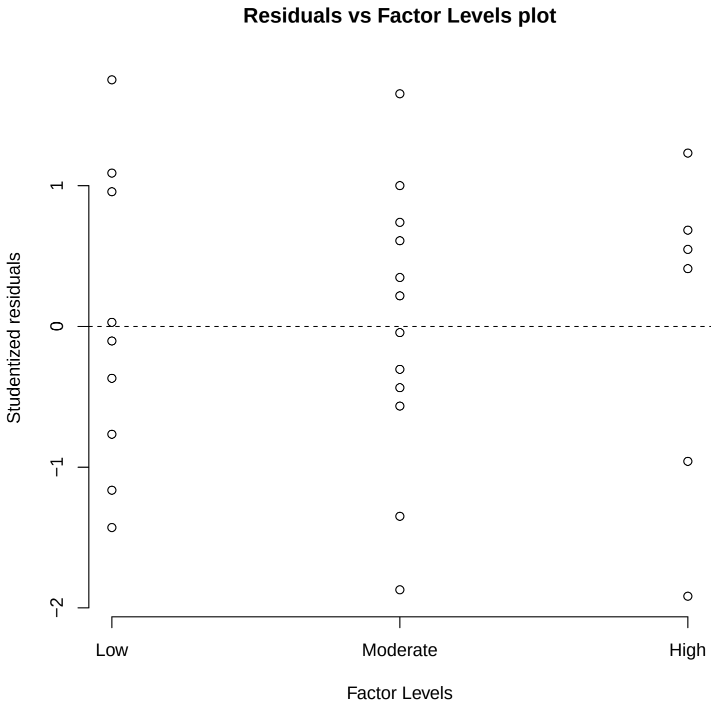
All this graphical instruments should be used with the idea that the standardized residuals should follow a symmetric distribution around 0, with not too heavy tails, under the estimated ANOVA model. No serial dependence between the observed residuals should be present under the assumed model. Let us go back to our example:

```
> ### Analysis of residuals
> ##  Residuals
>
> n_groups = c(rep(n_i[1],sum(data$group==1)),
+              rep(n_i[2],sum(data$group==2)),
+              rep(n_i[3],sum(data$group==3)))
```

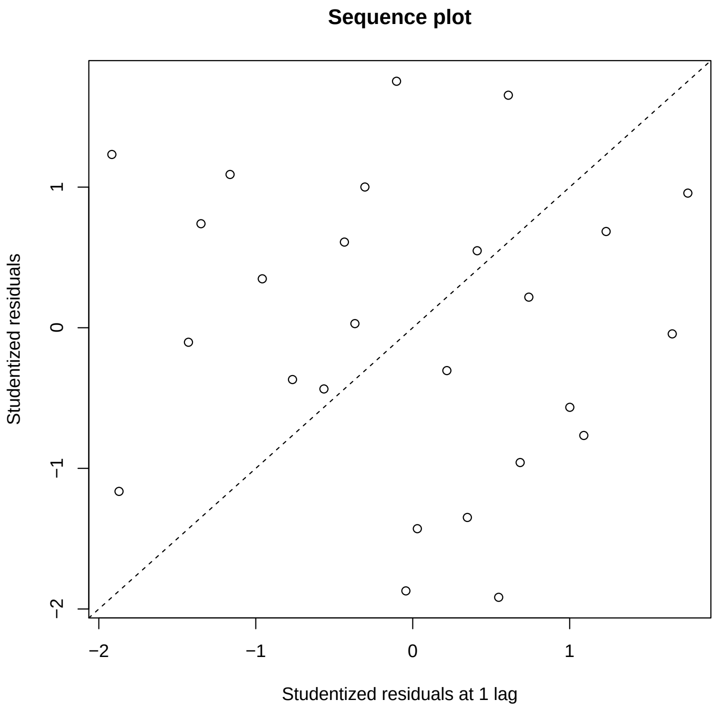
```
> e_star = residuals(mod)/sqrt(sse/(n_T-r))
> r_stud = residuals(mod)/sqrt((sse/(n_T-r))*(1-1/n_groups))
> t_deleted = residuals(mod)*sqrt((n_T-r-1)/(sse*(1-1/n_groups)-residuals(mod)^2))
> cbind("e"=data$y-fitted.values(mod),"e with R"=residuals(mod),"e star"=e_star,
+       "r stud"=r_stud,"r stud with R"=rstandard(mod),
+       "t deleted"=t_deleted,"t deleted with R"=rstudent(mod))
             e    e with R      e star      r stud r stud with R   t deleted t deleted with R
1   0.72222222  0.72222222  0.90271248  0.95747117    0.95747117  0.95574307       0.95574307
2   1.32222222  1.32222222  1.65265823  1.75290876    1.75290876  1.83766574       1.83766574
3  -0.07777778 -0.07777778 -0.09721519 -0.10311228   -0.10311228 -0.10096362      -0.10096362
4  -1.07777778 -1.07777778 -1.34712478 -1.42884160   -1.42884160 -1.46233725      -1.46233725
5   0.02222222  0.02222222  0.02777577  0.02946065    0.02946065  0.02884088       0.02884088
6  -0.27777778 -0.27777778 -0.34719711 -0.36825814   -0.36825814 -0.36152734      -0.36152734
7  -0.57777778 -0.57777778 -0.72216998 -0.76597694   -0.76597694 -0.75918652      -0.75918652
8   0.82222222  0.82222222  1.02770344  1.09004411    1.09004411  1.09453127       1.09453127
9  -0.87777778 -0.87777778 -1.09714286 -1.16369573   -1.16369573 -1.17276071      -1.17276071
10 -1.43333333 -1.43333333 -1.79153708 -1.87119945   -1.87119945 -1.98208267      -1.98208267
11 -0.03333333 -0.03333333 -0.04166365 -0.04351627   -0.04351627 -0.04260171      -0.04260171
12  1.26666667  1.26666667  1.58321881  1.65361812    1.65361812  1.71973234       1.71973234
13  ...............

> plot(fitted.values(mod),rstandard(mod),
+      xlab="Fitted values",
+      ylab="Studentized residuals",
+      main="Residuals vs fitted values plot")
> abline(h=0,lty="dashed")
```



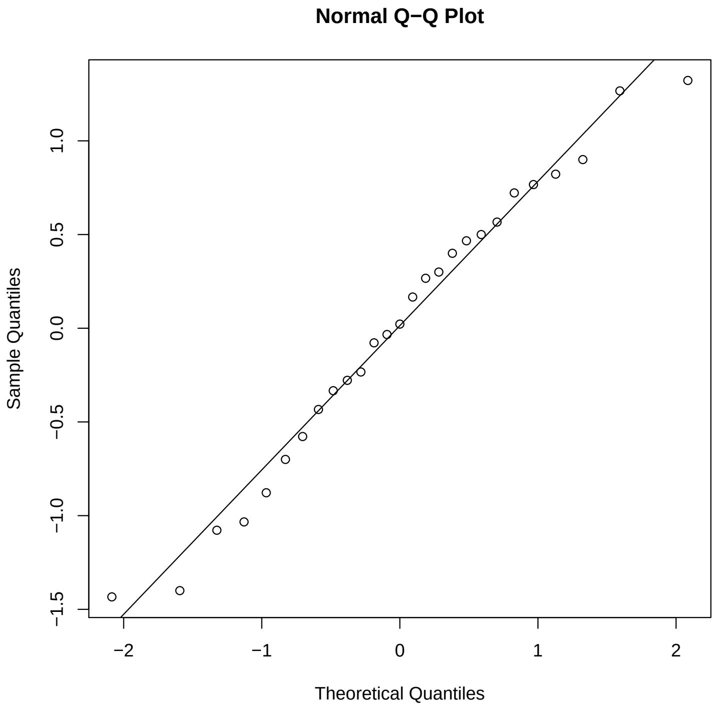
```
> plot(data$group,rstandard(mod),
+      xlab="Observed values",
+      ylab="Studentized residuals",
+       main="Residuals vs factor levels plot")
> abline(h=0,lty="dashed")
```



```
> plot(rstandard(mod)[-c(1)],rstandard(mod)[-c(n_T)],
+       xlab="Studentized residuals at 1 lag",
+       ylab="Studentized residuals",
+       main="Sequence plot")
> abline(a=0,b=1,lty="dashed")
```



```
qqnorm(residuals(mod))
qqline(residuals(mod))
```



1.2 Formal tests

Formal tests on residuals

After having explored the residuals by graphical means, we can also test more formally our hypotheses. We can derive formal tests to answer the following questions:

- is the variance of the error terms constant?
- are there outliers in the data?
- are the error terms normally distributed?

Tests for constancy of error variance

After having explored the residuals by graphical means, we can also test more formally our hypotheses. We can test the nonconstancy of the error variance with different procedures. We will describe the Hartley and the Brown-Forsythe tests. The Hartley test is indicated for balanced designs and when the errors are normally distributed. The Brown-Forsythe is instead more general, being robust to departures from normality and not needing equal sample sizes. The Hartley test compares the variance of r different normal populations, σ_i^2. It assumes that a random sample of constant sample size is drawn independently from each population. The tested hypotheses are

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 \\ H_1: & \text{at least one } \sigma_i^2 \text{ is different} \end{cases},$$

The test statistic is built using only the minimum and the maximum sample variances:

$$H^{obs} = \frac{max_{i=1,\dots,r}\left\{s_i^2\right\}}{min_{i=1,\dots,r}\left\{s_i^2\right\}},$$

where s_i^2 are the sample variances for the different populations.

The distribution of the test statistic H under the null hypothesis, $H^{obs} > H_{(r,df)}$, has been tabulated. It's shape changes according to the number of populations r and the degrees of freedom associated to the estimates s_i^2. Values of H^{obs} close to 1 are in favor of H_0, while we reject H_0 for large values of the test statistic (hence for $H^{obs} > H_{(1-\alpha;r,df)}$). In an ANOVA problem with balanced design ($n_i = n$, $i = 1, \dots, r$), the needed ingredients are the number of factor levels r, the degrees of freedom $df = n - 1$ and the estimates $s_i^2 = \sum_{j=1}^{n} e_{ij}^2 / \left(n-1\right)$. The Brown-Forsythe test compares the variance of r different normal populations, σ_i^2. It assumes that a random sample of possibly varying sample size is drawn independently from each population. Thus, the tested hypotheses are the same as for the Hartley test. The test statistic is built using the absolute deviations of the y_{ij} from their respective factor level medians $d_{ij} = |y_{ij} - \tilde{y}_i|$. The test is based on the fact that, under H_0, the expected values of the absolute deviations for the r factor levels are equal, while unequal variances σ_i^2 imply differing expected values for the absolute deviations.

Thus, it makes sense to construct a test statistic similar to the F test, but this time built on the deviations from the factor levels medians (instead of on the original observations):

$$F_{BF}^{obs} = \frac{MSTR_{BF}}{MSE_{BF}} = \frac{\frac{\sum_{i=1}^{r} n_i \left(\bar{d}_{i\cdot} - \bar{d}_{\cdot\cdot}\right)^2}{r-1}}{\frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(d_{ij} - \bar{d}_{i\cdot}\right)^2}{n_T - r}}.$$

The distribution of the test statistic F_{BF} under the null hypothesis is approximately $F_{(r-1,n_T-r)}$, provided that the sample sizes n_i are not too small. Values of F_{BF}^{obs} close to 1 are in favour of H_0, while we reject H_0 for large values of the test statistic (hence for $F_{BF}^{obs} > F_{(1-\alpha;r-1,n_T-r)}$).

Tests for outliers

After having identified possible outliers by graphical means, we can also test it more formally. We can test if a given residual is associated to an outlier observation by looking whether such residual falls in the tails of its null distribution. Hence, the idea is to check whether that residual is behaving strange with respect to its expected distribution under the assumed model. In practice, we need to compare the absolute value of the studentized deleted residuals with a critical value. In this case such value is $t_{(1-\alpha/(2n_T)),n_T-r-1)}$ (notice that the Bonferroni correction is included in this critical value).

Tests for normality

After having identified a possible departure form normality by graphical means, we can also test it more formally. We can test if the residuals follow a normal distribution by comparing the observed frequencies and the expected ones under normality. Shapiro-Wilk or Kolmogorov-Smirnov tests are a possible choice. Let us go back to our example:

```
> ### Formal tests
> ## Constancy of variances
> library(lawstat)
> levene.test(data$y,data$group1,location="median")

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  data$y
Test Statistic = 0.0245, p-value = 0.9758

>
> ## Outliers
> pvalue_outliers = NULL
> for(i in 1:n_T){
+ pvalue_outliers[i] = ((1-pt(abs(rstudent(mod)[i]),n_T-r-1))*2)*n_T
+ }
> pvalue_outliers[pvalue_outliers>1] = 1
> cbind("Stud  deleted res"=rstudent(mod),"Outlier p-value"=pvalue_outliers)
   Stud  deleted res Outlier p-value
1         0.95574307               1
2         1.83766574               1
3        -0.10096362               1
4        -1.46233725               1
5         0.02884088               1
6        -0.36152734               1
7        -0.75918652               1
8          .....

> ## Normality
> shapiro.test(residuals(mod))

        Shapiro-Wilk normality test

data:  residuals(mod)
W = 0.9738, p-value = 0.7033

> ks.test(residuals(mod),"pnorm",alternative="two.sided")

        One-sample Kolmogorov-Smirnov test
```

```
data:  residuals(mod)
D = 0.11, p-value = 0.8645
alternative hypothesis: two-sided
```

1.3 Remedial measures

Remedial measures

When the residual analysis shows departures from the assumed model, we need to adopt remedial measures. This practice consists of:

- modifying (in order to improve) the estimates of the model or
- transforming, somehow, the response Y, and repeat the standard estimation procedures.
- adopting a nonparametric test to make inference.

In particular we aim at solving:

- nonconstancy of the variance: in this case we can consider *weighted least squares* estimation techniques.
- nonnormality of the error terms: in this case we can consider a transformation of the response Y, to fit the ANOVA model on.
- larger departures from the assumed model: in this case we can consider a nonparametric alternative to the classical inferential procedures.

Weighted least squares

When the residual analysis shows that the variance is not constant over the factor levels, we can try to solve the problem by using *weighted least squares* estimation techniques. This way we can improve our estimates, and reach a much better fitting model. In practice, the model is fitted by weighting the observed y_{ij} with weights that are inverse proportional to the variance σ_i^2, associated to the several factor levels. Then, a modified version of the F test is adopted, which is based on the new sums of squares. This test can then be used to perform a more reliable inference on the data.

Transformation of Y

When the residual analysis shows a departure from the normal distribution and/or nonconstant error variance, we can try to transform the response Y, in order to get a error distribution which is closer to the normal one and constant in variance. Then, we can fit the ANOVA model on the new outcome, in order to obtain more reliable inferential conclusions. Some simple guides can be used in order to find the suitable transformation for general departures from the model assumptions. These guides can be used to overcome the nonconstancy of the error variance, but they are useful also to obtain a distribution of the error term closer to the normal one. We can distinguish several cases according to the relation between the error variance and the μ_i:

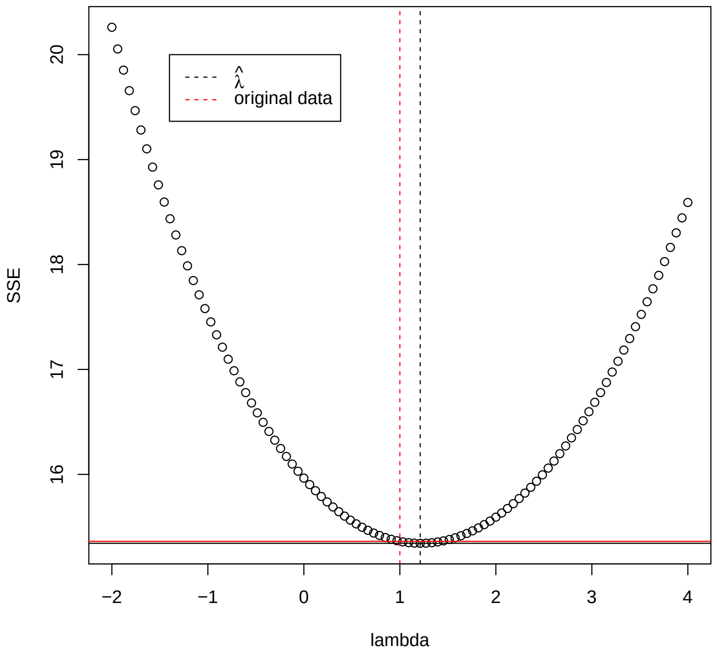
- σ_i^2 is proportional to μ_i (often for count data): take $Y^{''} = \sqrt{Y}$ or $Y^{''} = \sqrt{Y+1}$,
- σ_i is proportional to μ_i: take $Y^{'} = \log Y$,
- σ_i is proportional to μ_i^2: take $Y^{'} = 1/Y$,
- σ_i^2 is not constant because Y is a proportion: take $Y^{'} = 2 \arcsin \sqrt{Y}$.

These are just general guides. However, we can be more precise, case by case, using the so called *Box-Cox procedure*. The *Box-Cox procedure* identifies a power transformation of the type Y^{λ} to correct for both lack of normality and nonconstancy of the error variance. The procedure works as follows:

- a set of possible values for λ is chosen, $\{\lambda_s\}_{s\in S}$,
- for each $s \in S$, the observed data y is transformed according to $y^{'}_{ij} = y^{\lambda_s}_{ij}$. An ANOVA model is then fitted (after standardization of $y^{'}$), and the error sum of squares SSE_s is computed for each $s \in S$,
- the value of λ, $\hat{\lambda}$, that minimizes SSE is chosen as the best transformation.

Because SSE as a function of λ is usually quite flat in the neighborhood of its minimum, then a more meaningful value close to $\hat{\lambda}$ can be chosen. Let us go back to our example:

```
> ### Box-Cox procedure
> lambda = seq(-2,4,length=100)
> sse_s = NULL
> for(s in 1:length(lambda)){
+ y_new = (data$y)^lambda[s]
+ # we need to standardize y (see Kutner et al. (2005), Chapter 18)
+ K2 = (prod(data$y))^(1/n_T)
+ K1 = 1/(lambda[s]*(K2^(lambda[s]-1)))
+ if(lambda[s]!=0){
+ w_new = K1*(data$y^lambda[s]-1)}
+ if(lambda[s]==0){
+ w_new = K2*log(data$y)}
+ mod_new = lm(w_new~data$group1)
+ sse_s[s] = deviance(mod_new)
+ }
> which_lambda_hat = which.min(sse_s)
> lambda_hat = lambda[which_lambda_hat]
> lambda_hat
[1] 1.212121

> plot(lambda,sse_s,
+      xlab="lambda",ylab="SSE")
> abline(h=min(sse_s))
> abline(v=lambda_hat,lty="dashed")
> abline(h=deviance(mod),col=2)
> abline(v=1,lty="dashed",col=2)
> legend(-1.4,20,lty=c(2,2),col=c(1,2),legend=c(expression(hat(lambda)),"original data"))
```



Nonparametric rank F test

When the residual analysis shows a departure from the normal distribution, we can use a nonparametric version of the F test. The idea is to use a testing procedure which is not based on the assumption of normality of the error terms. Then the null hypothesis to test becomes that the r populations under study are continuous distributions having the same, unknown, distribution. Under the alternative hypothesis, the r populations are assumed to have distributions that differ only in the location (mean or median). This problem can be solved by applying a rank F^* test. The F^* test works according to the following steps:

- the observations y_{ij} are transformed in ranks (in case of ties, the mean rank is assigned to the tied observations),
- the standard F statistic is applied to the ranks,
- and it is compared with its approximate null distribution $F_{(r-1,n_T-r)}$ (provided that the sample sizes n_i are not too small).

It is also possible to derive partial tests for the single comparisons of factor levels means based on rank data, and using the Bonferroni correction.

1.4 Effects of departures from the model

Effects of departure from the model

We just discussed how to deal with some departures from the model (nonconstancy of the error variance and nonnormality of the errors). To adopt some of the techniques we saw, can help us with reducing the magnitude of the problem, but afterwards there might still be some assumptions which are not met yet. The effects of these departures from the model can still be not acceptable, and must be taken into consideration: if such departures are too extreme, then other statistical techniques should be taken into consideration. These are some effects of departures from the model:

- *Nonnormality*: lack of normality is not a too severe matter, provided that it not too extreme. In particular, kurtosis of the error distribution is affecting inference more than skewness. In general, the factor level means estimators are still unbiased in case of nonnormality, but the F test's type I error can be affected by this distributional departure (actual α of the test can be larger than the nominal one). Such influence on the inferential error of the F tests is, however, quite contained.

- *Nonconstancy of error variance*: this departure from the model is not affecting too badly the F test conclusions when the sample sizes for the factor levels are equal or similar. This is not the case, though, for the single factor level means comparisons, for which the results become not reliable even in case of equal sample sizes.

- *Nonindependence of the error terms*: it can have serious effects on inference. It is very important to remove the dependence, in order to have reliable results.

Summarizing (and more), some hints to decide what to do:

Departure	**Effect**	**Solution**
Nonnormality	Typically not too bad for F test and Sheffé contrasts estimates	Use transformations of Y. Use nonparametric F test. Use Generalized Linear Model (GLM).
Nonconstancy of the variance	Typically not too bad for F test and Sheffé contrasts estimates in balanced designs. In general bad effect on single factor levels means comparisons	Use transformations of Y. Use weighted least squares. Use Generalized Linear Model (GLM).
Nonindependence	Typically very bad	Use time series model.
Outliers	Results can change	Discard outliers (be very careful). Add interaction with other covariates. Use robust estimation.

ANALYSIS OF VARIANCE
Master of Statistics

Lesson 4

Dr. Francesca Solmi

November 12th, 2013

Contents

1 Model reformulation and regression approach

1.1 The linear regression model

The regression model

Kutner et al. (2005), Chapter 6 - Section 8. A regression model is used to explain a dependent variable Y in terms of one or more independent variables $\boldsymbol{X}^{'} = (X_1, \dots, X_p)$ (reminds of the starting point for ANOVA model). When, in general, Y is a quantitative random variable, and $\boldsymbol{X}^{'}$ can take both quantitative and qualitative values, then we can describe a *regression model* as

$$Y = f(X) + \epsilon$$

where we are approximating Y with a certain function $f(\cdot)$ of $\boldsymbol{X}^{'}$ (the approximation is implied by the presence of the error ϵ). We can then specify a *linear* regression model by imposing a linear shape to $f(\cdot)$. Hence

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \epsilon$$

We saw in the first lesson that there is a strong connection between a regression and an ANOVA model. Indeed, we can rewrite any ANOVA model in terms of regression model, by defining suitable covariates $\boldsymbol{X}^{'}$. We could see the relation between the ANOVA and regression models already through all the previous lessons, when in the R example we used the function `lm()`, which estimates a linear regression model on the data. Now we will focus on three different ways of representing an ANOVA model, and show how they can be represented as regression models, after a suitable choice of the covariates. These three models are:

- *Factor effect model with unweighted mean*:

$$y_{ij} = \mu. + \tau_i + \epsilon_{ij}, \quad \text{where } \mu. \text{ is the average of the factor level means.}$$

- *Factor effect model with weighted mean*:

$$y_{ij} = \mu. + \tau_i + \epsilon_{ij}, \quad \text{where } \mu. \text{ is a weighted average of the factor level means.}$$

- *Cell means model*:

$$y_{ij} = \mu_i + \epsilon_{ij}.$$

It is important to underline that the different representations of the ANOVA model as linear regression model affect only the definition of the parameters, and NOT the inference that can be done within the linear regression framework to test for the difference between the factor levels means.

1.2 Factor effects model

Model formulation

The *factor effect model* is an alternative way of formulating the cell mean model. It can be represented as:

$$y_{ij} = \mu. + \tau_i + \epsilon_{ij},$$

where

- $\mu.$ is a constant component common to all observations,
- τ_i is the effect of the i^{th} factor level,
- ϵ_{ij} are, again, independent realizations from a $N(0, \sigma^2)$.

According to how the general mean $\mu.$ is defined, we can distinguish between *factor effect model with unweighted or weighted mean.* **Unweighted mean** Here $\mu.$ is the unweighted mean of the factor level means $\mu. + \tau_i$. This means that a constraint is implied on the τ_i parameters:

$$\sum_{i=1}^{r} \tau_i = 0.$$

This implies that the parameters we need to estimate in the model are $\mu.$ and $r-1$ of the τ_i, since one of them can be expressed in terms of the others. For instance, we can drop the parameter $\tau_r = -\tau_1 - \tau_2 - \cdots - \tau_{r-1}$. We can now develop the corresponding linear model. Let us consider, as an example, a single-case study with $r = 3$ factor levels and $n_1 = n_2 = n_3 = 3$. Then the model can be written as

$$\overbrace{\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}}^{\boldsymbol{Y}} = \overbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix}}^{\boldsymbol{X}} \overbrace{\begin{pmatrix} \mu. \\ \tau_1 \\ \tau_2 \end{pmatrix}}^{\boldsymbol{\beta}} + \overbrace{\begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{pmatrix}}^{\boldsymbol{\epsilon}} = \begin{pmatrix} \mu. + \tau_1 + \epsilon_{11} \\ \mu. + \tau_1 + \epsilon_{12} \\ \mu. + \tau_1 + \epsilon_{13} \\ \mu. + \tau_2 + \epsilon_{21} \\ \mu. + \tau_2 + \epsilon_{22} \\ \mu. + \tau_2 + \epsilon_{23} \\ \mu. - \tau_1 - \tau_2 + \epsilon_{31} \\ \mu. - \tau_1 - \tau_2 + \epsilon_{32} \\ \mu. - \tau_1 - \tau_2 + \epsilon_{33} \end{pmatrix}$$

Notice that the covariates are not dummy variables but they can take values -1, 0 and 1. Thus, in general for a single-factor study with r factor levels, the multiple regression model can be written as

$$y_{ij} = \mu. + \tau_1 x_{ij,1} + \tau_2 x_{ij,2} + \cdots + \tau_{r-1} x_{ij,r-1} + \epsilon_{ij},$$

where the covariates are defined as

$$x_{ij,1} = \begin{cases} 1 & \text{if case from factor level 1} \\ -1 & \text{if case from factor level } r \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$x_{ij,r-1} = \begin{cases} 1 & \text{if case from factor level } r-1 \\ -1 & \text{if case from factor level } r \\ 0 & \text{otherwise} \end{cases} .$$

Hence, in terms of regression model, $\mu.$ is the intercept term, and $\tau_1, \ldots, \tau_{r-1}$ are the regression parameters. The least squares estimates for the parameters are (notice that in general $\hat{\mu}. \neq \bar{y}..$):

$$\hat{\mu}. = \frac{\sum_{i=1}^{r} \bar{y}_{i\cdot}}{r} \quad \text{and} \quad \hat{\tau}_i = \bar{y}_{i\cdot} - \hat{\mu}..$$

We can test the equality of the factor levels means in terms of the regression parameters:

$$\begin{cases} H_0 : & \tau_1 = \tau_2 = \cdots = \tau_{r-1} = 0 \\ H_1 : & \text{not all } \tau_i \text{ are equal to } 0 \end{cases} ,$$

hence by testing the nullity of all the regression coefficients. This testing problem can be solved by performing the corresponding F test to compare the full factor effects model with the reduced model $y_{ij} = \mu. + \epsilon_{ij}$. The test statistic is defined as

$$F^{regr} = \frac{MSR}{MSE},$$

where $SSR = MSR * df_R$ is the *regression sum of squares*, which coincides in this case with $SSTR$. The test statistic can be compared with its distribution under H_0, which is a $F_{(p-1,n-p)}$ *in regression notation*, where p is the number of parameters in the model (r in our case, $\mu.$ plus the $r-1$ τ_i) and n is the total sample size (n_T in our notation). Thus, the same conclusions are obtained following the regression approach.

Weighted mean Here $\mu.$ is the weighted mean of the factor level means $\mu. + \tau_i$. We could be interested in this kind of model formulation when:

- we have different sample sizes for the several factor levels, and we want to give more weight to factor levels containing more information,
- or it is meaningful to assign weights to the factor levels because, for example in prediction problems, we are interested in a general mean to which the several factor levels contribute with different importance.

In this case the weighted mean $\mu.$ is given by

$$\mu. = \sum_{i=1}^{r} w_i \mu_i, \quad \text{with} \sum_{i=1}^{r} w_i = 1$$

This means that, since

$$\mu. = \sum_{i=1}^{r} w_i \mu_i = \sum_{i=1}^{r} w_i \left(\mu. + \tau_i\right) = \mu. + \sum_{i=1}^{r} w_i \tau_i$$

then the constraint implied on the τ_i parameters is now:

$$\sum_{i=1}^{r} w_i \tau_i = 0$$

Again, this implies that the parameters we need to estimate in the model are $\mu.$ and $r-1$ of the τ_i, since one of them can be expressed in terms of the others. For instance, we can drop the parameters $\tau_r = -\frac{w_1}{w_r}\tau_1 - \frac{w_2}{w_r}\tau_2 - \cdots - \frac{w_{r-1}}{w_r}\tau_{r-1}$. Let us consider, for instance, the case of weights w_i proportional to the sample sizes n_i. In this case

$$w_i = \frac{n_i}{n_T}.$$

We now ready to develop the corresponding linear model. Let us consider again the example of a single-case study with $r = 3$ factor levels and $n_1 = n_2 = n_3 = 3$. Then the model can be written as

$$\overbrace{\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}}^{\boldsymbol{Y}} = \overbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -n_1/n_r & -n_2/n_r \\ 1 & -n_1/n_r & -n_2/n_r \\ 1 & -n_1/n_r & -n_2/n_r \end{pmatrix}}^{\boldsymbol{X}} \overbrace{\begin{pmatrix} \mu. \\ \tau_1 \\ \tau_2 \end{pmatrix}}^{\boldsymbol{\beta}} + \overbrace{\begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{pmatrix}}^{\boldsymbol{\epsilon}}$$

$$= \begin{pmatrix} \mu. + \tau_1 + \epsilon_{11} \\ \mu. + \tau_1 + \epsilon_{12} \\ \mu. + \tau_1 + \epsilon_{13} \\ \mu. + \tau_2 + \epsilon_{21} \\ \mu. + \tau_2 + \epsilon_{22} \\ \mu. + \tau_2 + \epsilon_{23} \\ \mu. - n_1/n_r\tau_1 - n_2/n_r\tau_2 + \epsilon_{31} \\ \mu. - n_1/n_r\tau_1 - n_2/n_r\tau_2 + \epsilon_{32} \\ \mu. - n_1/n_r\tau_1 - n_2/n_r\tau_2 + \epsilon_{33} \end{pmatrix}$$

Notice that, again the covariates are not dummy variables but they can take values $-n_1/n_r$, $-n_2/n_r$, 0 and 1. Thus, in general for a single-factor study with r factor levels, the multiple regression model can be written as

$$y_{ij} = \mu. + \tau_1 x_{ij,1} + \tau_2 x_{ij,2} + \cdots + \tau_{r-1} x_{ij,r-1} + \epsilon_{ij},$$

where the covariates are defined as

$$x_{ij,1} = \begin{cases} 1 & \text{if case from factor level 1} \\ -n_1/n_r & \text{if case from factor level } r \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$x_{ij,r-1} = \begin{cases} 1 & \text{if case from factor level } r-1 \\ -n_{r-1}/n_r & \text{if case from factor level } r \\ 0 & \text{otherwise} \end{cases} .$$

Hence, in terms of regression model, $\mu.$ is the intercept term, and $\tau_1, \ldots, \tau_{r-1}$ are the regression parameters. The least squares estimates for the parameters are in this case:

$$\hat{\mu}. = \sum_{i=1}^{r} \frac{n_i}{n_T} \bar{y}_{i}. = \bar{y}.. \quad \text{and} \quad \hat{\tau}_i = \bar{y}_i. - \hat{\mu}..$$

Again, we can test the equality of the factor levels means in terms of the regression parameters:

$$\begin{cases} H_0: & \tau_1 = \tau_2 = \cdots = \tau_{r-1} = 0 \\ H_1: & \text{not all } \tau_i \text{ are equal to } 0 \end{cases} ,$$

hence by testing the nullity of all the regression coefficients. Again, the same conclusions of the standard ANOVA test are obtained following the regression approach.

1.3 Cell means model

Model formulation

The *cell means model* is represented as:

$$y_{ij} = \mu_i + \epsilon_{ij}.$$

Hence, the parameters we need to estimate in the model are the μ_i: again, as for the factor effects model, we have r parameters in total in the model. Let us develop the corresponding linear model. Let us consider again the example of a single-case study with $r = 3$ factor levels and $n_1 = n_2 = n_3 = 3$. Then the model can be written as

$$\overbrace{\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}}^{\boldsymbol{Y}} = \overbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}}^{\boldsymbol{X}} \overbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}}^{\boldsymbol{\beta}} + \overbrace{\begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{pmatrix}}^{\boldsymbol{\epsilon}} = \begin{pmatrix} \mu_1 + \epsilon_{11} \\ \mu_1 + \epsilon_{12} \\ \mu_1 + \epsilon_{13} \\ \mu_2 + \epsilon_{21} \\ \mu_2 + \epsilon_{22} \\ \mu_2 + \epsilon_{23} \\ \mu_3 + \epsilon_{31} \\ \mu_3 + \epsilon_{32} \\ \mu_3 + \epsilon_{33} \end{pmatrix}$$

Notice that now there is no intercept term in the model. Thus, in general for a single-factor study with r factor levels, the multiple regression model can be written as

$$y_{ij} = \mu_1 x_{ij,1} + \mu_2 x_{ij,2} + \cdots + \mu_r x_{ij,r} + \epsilon_{ij},$$

where the covariates are defined as (dummy variables)

$$\tau_1 x_{ij,1} = \begin{cases} 1 & \text{if case from factor level 1} \\ 0 & \text{otherwise} \end{cases}$$
$$\vdots$$
$$\tau_1 x_{ij,r} = \begin{cases} 1 & \text{if case from factor level } r \\ 0 & \text{otherwise} \end{cases} .$$

Hence, now there is no intercept term, and there are r regression parameters $\mu_1, \ldots, \mu_r$. The least squares estimates for the parameters are in this case $\hat{\mu}_i = \bar{y}_{i\cdot}$. Now we can test the equality of the factor levels means in terms of the regression parameters:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_r \\ H_1: & \text{not all } \mu_i \text{ are equal} \end{cases},$$

hence this time we are NOT interested in testing the nullity of all the regression coefficients, BUT their equality. This testing problem can be solved by performing the corresponding F test to compare the full cell mean model with the reduced model $y_{ij} = \mu_c + \epsilon_{ij}$, where μ_c is the common mean when H_0 holds. The test statistic is defined as

$$F^{regr} = \frac{\overbrace{SSE_r - SSE_f}^{SSTO-SSE=SSR}}{\underbrace{df_r - df_f}_{(n-1)-(n-p)=p-1}} \Big/ \frac{\overbrace{SSE_f}^{SSE}}{\underbrace{df_f}_{n-p}} = \frac{SSR}{p-1} \Big/ \frac{SSE}{n-p} = \frac{MSR}{MSE},$$

so, again, the same conclusions are obtained following the regression approach.

1.4 R codes

How to do it with R

We already noticed that ANOVA problems can be dealt with in R using the `lm()` function. We just saw how several different parametrization of the model are possible. We will now see how to do it in R, using the functions `lm()` and `contrasts()`. Let us go back to our Productivity Improvement example:

```
> #### Regression approach
> ### How to define the contrasts in R
> ## Model matrixes per type of contrast
> model.matrix(~ group1, data);
   (Intercept) group12 group13
1            1       0       0
2            1       0       0
...
8            1       0       0
9            1       0       0
10           1       1       0
11           1       1       0
...
20           1       1       0
21           1       1       0
22           1       0       1
23           1       0       1
...
26           1       0       1
27           1       0       1
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$group1
[1] "contr.treatment"

> model.matrix(~ -1+group1, data)
   group11 group12 group13
1        1       0       0
2        1       0       0
...
8        1       0       0
9        1       0       0
10       0       1       0
11       0       1       0
...
20       0       1       0
21       0       1       0
22       0       0       1
23       0       0       1
...
26       0       0       1
27       0       0       1
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$group1
[1] "contr.treatment"

> model.matrix(~ group1, data, contrasts = list(group1="contr.treatment"))
   (Intercept) group12 group13
```

```
1            1       0       0
2            1       0       0
...
8            1       0       0
9            1       0       0
10           1       1       0
11           1       1       0
...
20           1       1       0
21           1       1       0
22           1       0       1
23           1       0       1
...
26           1       0       1
27           1       0       1
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$group1
[1] "contr.treatment"
```

```
> model.matrix(~ group1, data, contrasts = list(group1="contr.sum"))
   (Intercept) group11 group12
1            1       1       0
2            1       1       0
...
8            1       1       0
9            1       1       0
10           1       0       1
11           1       0       1
...
20           1       0       1
21           1       0       1
22           1      -1      -1
23           1      -1      -1
...
26           1      -1      -1
27           1      -1      -1
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$group1
[1] "contr.sum"
```

```
> ## Contrast matrixes per type of contrast
> contrasts(data$group1) = contr.treatment
> contrasts(data$group1)
  2 3
1 0 0
2 1 0
3 0 1
> contrasts(data$group1) = contr.sum
> contrasts(data$group1)
  [,1] [,2]
1    1    0
2    0    1
3   -1   -1
```

```
> C = matrix(c(1,0,-(n_i[1]/n_i[3]),0,1,-(n_i[2]/n_i[3])),byrow=F,nrow=3)
> contrasts(data$group1) = C
> model.matrix(~ group1, data)
   (Intercept) group11 group12
1            1     1.0       0
2            1     1.0       0
...
8            1     1.0       0
9            1     1.0       0
10           1     0.0       1
11           1     0.0       1
...
20           1     0.0       1
21           1     0.0       1
22           1    -1.5      -2
23           1    -1.5      -2
...
26           1    -1.5      -2
27           1    -1.5      -2
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$group1
  [,1] [,2]
1  1.0    0
2  0.0    1
3 -1.5   -2

> ###  Factor effects model
> ## Unweighted mean
> # define the contrasts
> contrasts(data$group1) = contr.sum
> # fit the model
> mod_fem.um = lm(y~group1,data=data)
> summary(mod_fem.um)

Call:
lm(formula = y ~ group1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43333 -0.50556  0.02222  0.53333  1.32222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.07037    0.16026  50.359  < 2e-16 ***
group11     -1.19259    0.22224  -5.366 1.65e-05 ***
group12      0.06296    0.20848   0.302    0.765
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8001 on 24 degrees of freedom
Multiple R-squared: 0.5671,     Adjusted R-squared: 0.531
F-statistic: 15.72 on 2 and 24 DF,  p-value: 4.331e-05

> data.frame("lm function"=c(coef(mod_fem.um)[1],coef(mod_fem.um)[1]+coef(mod_fem.um)[2],
+         coef(mod_fem.um)[1]+coef(mod_fem.um)[3],
+         coef(mod_fem.um)[1]-coef(mod_fem.um)[2]-coef(mod_fem.um)[3]),
+         "by hand"=c(mean(mu_i),mu_i[1],mu_i[2],mu_i[3]),row.names=c("mu","mu_1","mu_2","mu_3"))
     lm.function  by.hand
mu      8.070370 8.070370
mu_1    6.877778 6.877778
mu_2    8.133333 8.133333
mu_3    9.200000 9.200000
> # perform the anova test
> mod_H0 = lm(y~1,data=data)    # model under H_0
> anova(mod_fem.um,mod_H0)
Analysis of Variance Table

Model 1: y ~ group1
Model 2: y ~ 1
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     24 15.362
2     26 35.487 -2   -20.125 15.72 4.331e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ## Weighted mean
> # define the contrasts
> C = matrix(c(1,0,-(n_i[1]/n_i[3]),0,1,-(n_i[2]/n_i[3])),byrow=F,nrow=3)
> contrasts(data$group1) = C
> # fit the model
> mod_fem.wm = lm(y~group1,data=data)
> summary(mod_fem.wm)

Call:
lm(formula = y ~ group1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43333 -0.50556  0.02222  0.53333  1.32222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.9519     0.1540  51.645  < 2e-16 ***
group11      -1.0741     0.2177  -4.933 4.93e-05 ***
group12       0.1815     0.1721   1.054    0.302
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8001 on 24 degrees of freedom
Multiple R-squared: 0.5671,     Adjusted R-squared: 0.531
F-statistic: 15.72 on 2 and 24 DF,  p-value: 4.331e-05

> data.frame("lm function"=c(coef(mod_fem.wm)[1],coef(mod_fem.wm)[1]+coef(mod_fem.wm)[2],
+         coef(mod_fem.wm)[1]+coef(mod_fem.wm)[3],
+         coef(mod_fem.wm)[1]-coef(mod_fem.wm)[2]-coef(mod_fem.wm)[3]),
+         "by hand"=c(mean(data$y),mu_i[1],mu_i[2],mu_i[3]),row.names=c("mu","mu_1","mu_2","mu_3"))
     lm.function  by.hand
mu      7.951852 7.951852
mu_1    6.877778 6.877778
mu_2    8.133333 8.133333
```

```
mu_3    8.844444 9.200000
> # perform the anova test
> anova(mod_fem.wm,mod_H0)
Analysis of Variance Table

Model 1: y ~ group1
Model 2: y ~ 1
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     24 15.362
2     26 35.487 -2   -20.125 15.72 4.331e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> ###  Cell means model
> # define the contrasts
> contrasts(data$group1) = contr.treatment
> # fit the model
> mod_cmm = lm(y~-1+group1,data=data)
> summary(mod_cmm)

Call:
lm(formula = y ~ -1 + group1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43333 -0.50556  0.02222  0.53333  1.32222

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
group11   6.8778     0.2667   25.79   <2e-16 ***
group12   8.1333     0.2310   35.22   <2e-16 ***
group13   9.2000     0.3266   28.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8001 on 24 degrees of freedom
Multiple R-squared: 0.9912,     Adjusted R-squared: 0.9901
F-statistic: 899.6 on 3 and 24 DF,  p-value: < 2.2e-16
```

```
> data.frame("lm function"=c(coef(mod_cmm)[1],coef(mod_cmm)[2],coef(mod_cmm)[3]),
+         "by hand"=c(mu_i[1],mu_i[2],mu_i[3]),row.names=c("mu_1","mu_2","mu_3"))
     lm.function  by.hand
mu_1    6.877778 6.877778
mu_2    8.133333 8.133333
mu_3    9.200000 9.200000
> # perform the anova test
> anova(mod_cmm,mod_H0)
Analysis of Variance Table

Model 1: y ~ -1 + group1
Model 2: y ~ 1
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     24 15.362
2     26 35.487 -2   -20.125 15.72 4.331e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANALYSIS OF VARIANCE
Master of Statistics

Lessons 5 and 6

Dr. Francesca Solmi

Contents

1 Two-way ANOVA (Equal sample size)

1.1 Two factor Analysis of Variance

Setting up the scene
Kutner et al. (2005), Chapter 19.

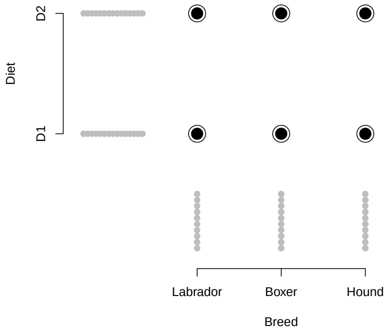
So far, we focused on single-factor studies. Now we go a step further into the study of the simultaneous effect of two or more factors. For now, we consider the case of two-factor studies, where the factors are crossed and all the sample sizes are equal. As single-factor studies, two-factor studies can be based on both experimental or observational studies. We will first present the problem through a simple example, which we will use during the entire lesson. Assume an experimental study is conducted about the body fat mass in dogs. Two factors are considered in the study:

- the diet given to the dogs (*Diet* $D1$ and *Diet* $D2$),
- the *breed* of the dog (three breeds are considered).

Of course, the two factors need to be well defined: the three breeds are chosen all among medium size: Labrador Retriver, Boxer and Afghan Hound are considered. The two diets are a low fat and a standard diet produced by a specific company. The company is interested in studying the difference of the two diets effect on the body fat, possibly distinguishing among the three breeds. All other factors that possibly influence the body fat are kept constant in the study. All dogs conduct the same life style, which includes physical activity for 1.5 hours and 6 hours of fresh air and socialization per day. Only females are considered in the study. Moreover the two diets are given to all the dogs from the age of 8 weeks, and the dogs body fat mass is measured at a specific age, the same for every single dog. The problem can be summarized as follows:

- we observe a quantitative response (body fat mass),
- there are two factors, namely A is the life style of the dog and B is the breed of the dog,
- factor A is studied at $a = 2$ levels, and factor B has $b = 3$ levels,
- each combination of factors A and B is a *treatment*, hence here we have $ab = 6$ treatments,

Let us assume that 10 dogs per breed are randomly allocated to the two diets with equal sample sizes, hence $n = 5$ dogs per group. The design layout can be displayed as follows:



Our objectives

We are interested in studying the effect of the two factors. We distinguish between main effects of the factors and their interactions. It is very important to work simultaneously with the two factors, otherwise we might end up with *wrong* conclusions. For instance, the difference in the effect of the two diets might be larger for Labradors than for Hounds, or the fat mass in Labradors and Boxers could be the same for dogs following diet $D1$ but it might be quite different for dogs that follow diet $D2$.

1.2 Meaning of ANOVA model elements

Assuming to know the real state of nature

We will now go through the ANOVA model considering our dogs' example. We will assume to know the real state of nature, and in particular the real means for the 6 treatments. We will face different possible situations, which will allow us to better understand the concepts of *main* factor effects and their *interactions*. We will denote the mean response (body fat mass) for a given treatment by μ_{ij}, where i refers to the level of factor A ($i = 1, \ldots, a$), and j refers to the level of factor B ($j = 1, \ldots, b$)

Additive factor effects - example 1

Let us assume that the real mean responses are (measured in fat percentage):

	Factor B - Breed			
	***j* = 1**	***j* = 2**	***j* = 3**	**Row**
Factor A - Diet	**Labrador**	**Boxer**	**Hound**	**average**
***i* = 1 D1**	24 (μ_{11})	22 (μ_{12})	20 (μ_{13})	22 ($\mu_{1\cdot}$)
***i* = 2 D2**	24 (μ_{21})	22 (μ_{22})	20 (μ_{23})	22 ($\mu_{2\cdot}$)
Column average	24 ($\mu_{\cdot 1}$)	22 ($\mu_{\cdot 2}$)	20 ($\mu_{\cdot 3}$)	22 ($\mu_{\cdot\cdot}$)

where the averages are given by $\mu_{\cdot j} = \frac{\sum_{i=1}^{a} \mu_{ij}}{a}$, $\mu_{i\cdot} = \frac{\sum_{j=1}^{b} \mu_{ij}}{b}$ and $\mu_{\cdot\cdot} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} \mu_{ij}}{ab} = \frac{\sum_{i=1}^{a} \mu_{i\cdot}}{a} = \frac{\sum_{j=1}^{b} \mu_{\cdot j}}{b}$. The main effects can be computed as:

Main Diet Effect	Main Breed Effect
$\alpha_1 = \mu_{1\cdot} - \mu_{\cdot\cdot} = 22 - 22 = 0$	$\beta_1 = \mu_{\cdot 1} - \mu_{\cdot\cdot} = 24 - 22 = 2$
$\alpha_2 = \mu_{2\cdot} - \mu_{\cdot\cdot} = 22 - 22 = 0$	$\beta_2 = \mu_{\cdot 2} - \mu_{\cdot\cdot} = 22 - 22 = 0$
	$\beta_3 = \mu_{\cdot 2} - \mu_{\cdot\cdot} = 20 - 22 = -2$

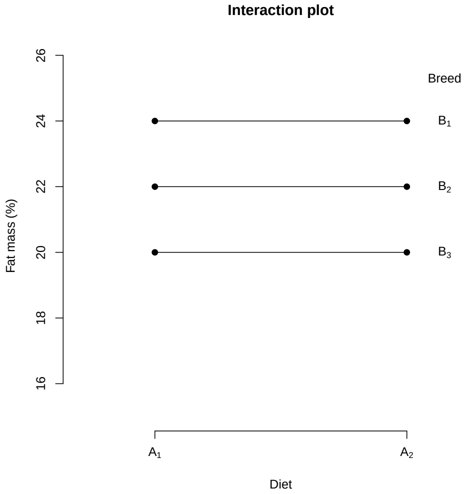
Thus, in this scenario, breed has an effect of the fat mass, while the diet does not. In general we define the main effects as

$$\alpha_i = \mu_{i\cdot} - \mu_{\cdot\cdot} \quad \text{and} \quad \beta_j = \mu_{\cdot j} - \mu_{\cdot\cdot}$$

The sum of the main effects for each factor is zero: $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$. We can notice that, in this case, each treatment mean can be obtained by summing the specific factor effects, thus

$$\begin{aligned} \mu_{ij} &= \overbrace{\mu_{\cdot\cdot} + \alpha_i + \beta_j}^{\text{in terms of the main effects}} \\ &= \overbrace{\mu_{i\cdot} + \mu_{\cdot j} - \mu_{\cdot\cdot}}^{\text{in terms of the row/column means}} \\ &= \overbrace{\mu_{ij'} + \mu_{i'j} - \mu_{i'j'}}^{\text{in terms of three other treatments means}} \quad , \quad i \neq i^{'}, j \neq j^{'} . \end{aligned}$$

When any of the treatments means can be written in one of the forms above, we say that *the factor effects are additive*, or, equivalently, that *no factor interactions are present*. This means that *the effect* *of either factor does not depend on the level of the other factor.* In our example the factors are additive, there is no interaction effect. In the specific case, the breed effect remains the same for all levels of the diet, thus the breed effect does not depend on the level of diet. This situation can be visualized in an *interaction plot.*

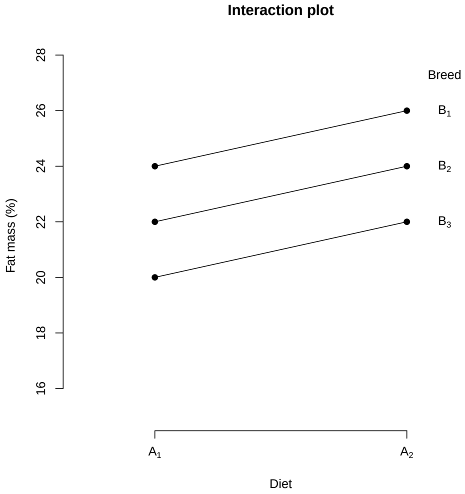


Additive factor effects - example 2

Let us assume that the real mean responses are (measured in fat percentage):

	Factor B - Breed			
	$j = 1$	$j = 2$	$j = 3$	**Row**
Factor A - Diet	**Labrador**	**Boxer**	**Hound**	**average**
$\boldsymbol{i=1}$ $D1$	24 (μ_{11})	22 (μ_{12})	20 (μ_{13})	22 $(\mu_{1\cdot})$
$\boldsymbol{i=2}$ $D2$	26 (μ_{21})	24 (μ_{22})	22 (μ_{23})	24 $(\mu_{2\cdot})$
Column average	25 $(\mu_{\cdot 1})$	23 $(\mu_{\cdot 2})$	21 $(\mu_{\cdot 3})$	23 $(\mu_{\cdot\cdot})$

The factor effects are still additive (check it!). The interaction plot becomes:



Interacting factor effects - example 3

Let us assume that the real mean responses are (measured in fat percentage):

	Factor B - Breed			
	$j = 1$	$j = 2$	$j = 3$	**Row**
Factor A - Diet	**Labrador**	**Boxer**	**Hound**	**average**
$i = 1$ $D1$	24 (μ_{11})	22 (μ_{12})	20 (μ_{13})	22 $(\mu_{1\cdot})$
$i = 2$ $D2$	28 (μ_{21})	24 (μ_{22})	20 (μ_{23})	24 $(\mu_{2\cdot})$
Column average	26 $(\mu_{\cdot 1})$	23 $(\mu_{\cdot 2})$	20 $(\mu_{\cdot 3})$	23 $(\mu_{\cdot\cdot})$

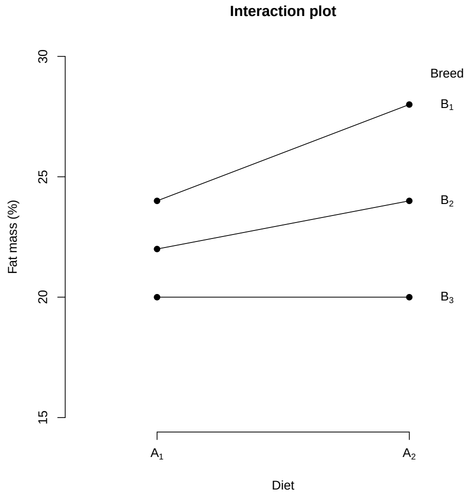
Notice that $\mu_{\cdot\cdot}$ is still the average of the row/column means. We can recognize the presence of the interaction by checking the additivity of the main factor effects. For instance

$$\mu_{11} \neq \mu_{\cdot\cdot} + \alpha_1 + \beta_1 = 23 + (22 - 23) + (26 - 23) = 23 - 1 + 3 = 25$$

The difference between μ_{11} and $\mu_{\cdot\cdot}+\alpha_1+\beta_1$ is called *interaction effect.* In general we define the *interaction* of the i^{th} level of factor A and the j^{th} level of factor B as

$$(\alpha\beta)_{ij} = \mu_{ij} - (\mu_{\cdot\cdot} + \alpha_i + \beta_j) = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}.$$

When the interaction effects are different from 0, we say that the factor effects are not additive, or, in other words, that *the effect of one factor depends on the level of the other factor.* In our last example the factors are not additive, thus there is interaction effect. In the specific case, the diet effect changes according to the breed. This situation can be visualized in the *interaction plot*: notice that this time the lines are not parallel.



1.3 ANOVA model definition

ANOVA model for two-factor studies

We can now develop the ANOVA model for two-factor studies *when all treatment sample sizes are equal (and all treatment means are of equal importance).* We will follow the same structure of lesson 1, hence we will:

- define the ANOVA model,

- present the estimation methods,
- define the concepts of $SSTO$, $SSTR$, SSE,
- present the F test.

Model definition

Let us suppose to deal with a two-factor study, where factor A has a levels and factor B has b levels. All ab factor combinations are included in the study. Let us also suppose that $n > 1$ observations of the response of interest Y are recorded for each of the factor combinations, and hence we deal with a total sample size of $n_T = abn$. We denote the single observation as y_{ijk}, where the indexes $i = 1, \ldots, a$ and $j = 1, \ldots, b$ refer to the factor levels, and the index $k = 1, \ldots, n$ denote the single observation for the specific treatment ij.

Cell Means model

We can express the ANOVA model in terms of the treatment means μ_{ij} as

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n,$$

where

- μ_{ij} are parameters,
- ϵ_{ijk} are independent realizations from $N(0, \sigma^2)$

Through some simple calculations we can show that (check why!):

- $E[Y_{ijk}] = \mu_{ij}$,
- $V[Y_{ijk}] = \sigma^2$, *constant*,
- Y_{ijk} are *normally distributed*,
- y_{ijk} are *independent* realizations from the corresponding distributions.

Factor Effects model

We can express the ANOVA model in terms of the factor effects α_i and β_j and their interactions $(\alpha\beta)_{ij}$ as

$$y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n,$$

where

- $\mu_{\cdot\cdot}$ is a parameter,
- α_i and β_j are parameters subject to restrictions $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$,
- $(\alpha\beta)_{ij}$ are parameters subject to the restriction

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 \quad j = 1, \ldots, b, \sum_{j=1}^{b} (\alpha\beta)_{ij} \qquad\qquad = 0 \quad i = 1, \ldots, a,$$

- ϵ_{ijk} are independent realizations from $N(0, \sigma^2)$.

We find here the same properties of the cell means model, just rewritten in terms of the new parameters:

- $E[Y_{ijk}] = \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$,
- $V[Y_{ijk}] = \sigma^2$, *constant*,
- Y_{ijk} are *normally distributed*,
- y_{ijk} are *independent* realizations from the corresponding distributions.

Estimation methods

The parameters of the ANOVA model need to be estimated. According to which is the model we are working with we will estimate directly μ_{ij} or α_i, β_j and $(\alpha\beta)_{ij}$. Least squares (LS) and the maximum likelihood (ML) estimates can be used, and once again they coincide for these models. Let us first introduce some notation:

- sum of the observations for the treatment ij: $y_{ij\cdot} = \sum_{k=1}^{n} y_{ijk}$,
- mean of the observations for the treatment ij: $\bar{y}_{ij\cdot} = \sum_{k=1}^{n} y_{ijk}/n$,
- sum of the all observations for the i^{th} level of factor A: $y_{i\cdot\cdot} = \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}$,
- mean of the all observations for the i^{th} level of factor A: $\bar{y}_{i\cdot\cdot} = \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}/(bn)$,
- sum of the all observations for the j^{th} level of factor B: $y_{\cdot j\cdot} = \sum_{i=1}^{a} \sum_{k=1}^{n} y_{ijk}$,
- mean of the all observations for the j^{th} level of factor B: $\bar{y}_{\cdot j\cdot} = \sum_{i=1}^{a} \sum_{k=1}^{n} y_{ijk}/(an)$,
- sum of the all observations in the study: $y_{\cdots} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}$,
- mean of the all observations in the study: $\bar{y}_{\cdots} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}/(abn)$.

Let us consider the LS method: we need to minimize the sum of squared deviations of the observations from their expected values:

$$\begin{aligned} Q =& \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(y_{ijk} - \mu_{ij}\right)^2 \\ =& \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(y_{ijk} - \mu_{\cdot\cdot} - \alpha_i - \beta_j - (\alpha\beta)_{ij}\right)^2 , \end{aligned}$$

according to which is the model formulation we choose (for the factor levels model, the minimization problem is subject to the constraints on α_i, β_j and $(\alpha\beta)_{ij}$). The solution is then given by:

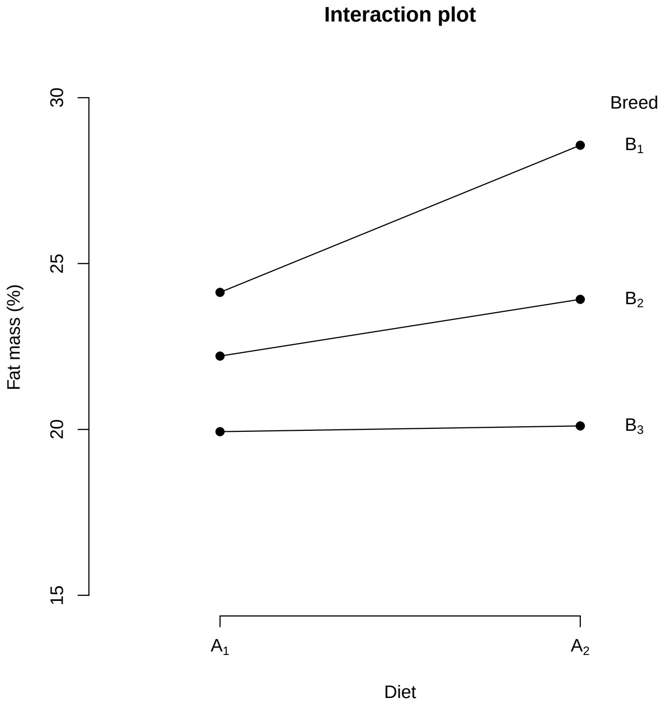
Parameter	Estimator
μ_{ij}	$\hat{\mu}_{ij} = \bar{Y}_{ij\cdot}$
$\mu_{\cdot\cdot}$	$\hat{\mu}_{\cdot\cdot} = \bar{Y}_{\cdots}$
$\alpha_i = \mu_{i\cdot} - \mu_{\cdot\cdot}$	$\hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}$
$\beta_j = \mu_{\cdot j} - \mu_{\cdot\cdot}$	$\hat{\beta}_j = \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots}$
$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot} + \mu_{\cdot\cdot}$	$(\hat{\alpha}\beta)_{ij} = \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots}$

We can also recover the *fitted values* $\hat{y}_{ijk} = \bar{y}_{ij\cdot}$ and the residuals $e_{ijk} = y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \bar{y}_{ij\cdot}$. Again, residuals can be used to examine the aptness of the model for a given data set. Let us assume now to observe the data from our experiment, and the real state of nature to be the last one considered (presence of interaction). We can simulate the data according to a cell means model:

```
> means_ex3 = array(c(24,28,22,24,20,20),dim=c(2,3))
> n = 10
> fat_11 = rnorm(n,means_ex3[1,1],sd=1)
> fat_12 = rnorm(n,means_ex3[1,2],sd=1)
> fat_13 = rnorm(n,means_ex3[1,3],sd=1)
> fat_21 = rnorm(n,means_ex3[2,1],sd=1)
> fat_22 = rnorm(n,means_ex3[2,2],sd=1)
> fat_23 = rnorm(n,means_ex3[2,3],sd=1)
> fat = c(fat_11,fat_12,fat_13,fat_21,fat_22,fat_23)
>
> hat_mu = array(c(mean(fat_11),mean(fat_21),
+                  mean(fat_12),mean(fat_22),
+                  mean(fat_13),mean(fat_23)),dim=c(2,3))
```

The interaction plot obtained from the estimates of the treatment means $\hat{\mu}_{ij}$ appears as:

```
> # Interaction plot
> plot(c(1,2),hat_mu[,1],main="Interaction plot",
+      xlab="Diet",ylab="Fat mass (%)",type="l",
+      axes=F,ylim=c(15,max(hat_mu[2,])+2),
+      xlim=c(0.7,2.3))
> points(c(1,2),hat_mu[,1],pch=19)
> points(c(1,2),hat_mu[,2],type="l")
> points(c(1,2),hat_mu[,2],pch=19)
> points(c(1,2),hat_mu[,3],type="l")
> points(c(1,2),hat_mu[,3],pch=19)
> axis(1,at=c(1,2),labels=c(expression(A[1]),
      expression(A[2])))
> axis(2)
> text(2+0.15,hat_mu[2,1]+1.3,label="Breed")
> text(2+0.15,hat_mu[2,1],label=expression(B[1]))
> text(2+0.15,hat_mu[2,2],label=expression(B[2]))
> text(2+0.15,hat_mu[2,3],label=expression(B[3]))
```



SSTO, *SSTR*, *SSE*

As for single-factor models, we can partition the total variability of the observations in difference parts:

- to distinguish among difference sources of variability and assigne to each of them their relative importance.

- to understand how much the fitted ANOVA model is able to improve the simplest possible model (which assumes an overall common mean for all the factors levels)

In practice:

- the total variability of the observations is measured in terms of the total deviation of the observations around the overall mean $(y_{ij} - \bar{y}_{\cdot\cdot})$.
- Once we fit the two-factor ANOVA model, we can decompose this total deviation in
  - first we view the study as a one factor study with ab levels (and we proceed as we saw in lesson 1):
    - the deviation of the observations from their specific factor levels estimates $(y_{ijk} - \bar{y}_{ij\cdot})$ and
    - the variability of the factor levels estimates from the overall mean $(\bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot\cdot})$.
  - then we further decompose the $\bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot\cdot}$ in terms of components that reflect the main effects of factors A and B and their interaction.

Thus we obtain:

$$\underbrace{y_{ijk} - \bar{y}_{\cdot\cdot\cdot}}_{\substack{\text{Total}\\ \text{deviation}}} = \underbrace{(y_{ijk} - \bar{y}_{ij\cdot})}_{\substack{\text{Deviation}\\ \text{around estimated}\\ \text{treatment mean}}} + \underbrace{(\bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot\cdot})}_{\substack{\text{Deviation of estimated}\\ \text{treatment mean around}\\ \text{overall mean}}} \quad ,$$

where

$$\underbrace{(\bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot\cdot})}_{\substack{\text{Deviation of estimated}\\ \text{treatment mean around}\\ \text{overall mean}}} = \underbrace{(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})}_{\substack{\text{A main}\\ \text{effect}}} + \underbrace{(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})}_{\substack{\text{B main}\\ \text{effect}}} + \underbrace{(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})}_{\substack{\text{AB interaction}\\ \text{effect}}}$$

If now we take the square of both sides of the equation, the cross-product terms drop out, and we obtain:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2}_{\text{Total sum of squares } (SSTO)} = \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij\cdot})^2}_{\text{Error sum of squares } (SSE)} + \underbrace{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot\cdot})^2}_{\text{Treatment sum of squares } (SSTR)} \quad ,$$

and

$$\underbrace{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot\cdot})^2}_{\text{Treatment sum of squares } (SSTR)} = \underbrace{nb\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2}_{\text{Factor } A \text{ sum of squares } (SSA)} + \underbrace{na\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2}_{\text{Factor } B \text{ sum of squares } (SSB)}$$

$$+ \underbrace{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})^2}_{AB \text{ interaction sum of squares } (SSAB)} .$$

Variance components' degrees of freedom

We can also easily obtain the degrees of freedom (*df*, number of independent/free observations for the estimation) associated to this variance decomposition

- $SSTO$ has $n_T - 1$ *df*.
- $SSTR$ has $ab - 1$ *df*, divided then in

  - SSA with $a-1$ df,
  - SSB with $b-1$ df,
  - $SSAB$ with $(a-1)(b-1)$ df.

- SSE has $(n-1)ab$ df.

Mean squares

The mean squares are obtained by dividing each sum of squares by it's associated df, thus

$$\underbrace{MSA = \frac{SSA}{a-1}}_{\text{Factor } A \text{ mean square}}, \quad \underbrace{MSB = \frac{SSB}{b-1}}_{\text{Factor } B \text{ mean square}} \quad \text{and} \quad \underbrace{MSAB = \frac{SSAB}{(a-1)(b-1)}}_{AB \text{ interaction mean square}}$$

These mean squares represent the average squared deviations, hence they are basically variance estimates (recall that $V[Z] = E[(Z - E[Z])^2]$).

The ANOVA table

To summarize:

Source of Variation	**SS**	**df**	**MS**	E[**MS**]
Factor A	$\overbrace{nb\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2}^{SSA}$	$a-1$	$\overbrace{\frac{SSA}{a-1}}^{MSA}$	$\sigma^2 + bn\frac{\sum_{i=1}^{a}(\mu_{i\cdot} - \mu_{\cdot\cdot})^2}{a-1}$
Factor B	$\overbrace{na\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2}^{SSB}$	$b-1$	$\overbrace{\frac{SSB}{b-1}}^{MSB}$	$\sigma^2 + an\frac{\sum_{j=1}^{b}(\mu_{\cdot j} - \mu_{\cdot\cdot})^2}{b-1}$
AB interaction	$\overbrace{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})^2}^{SSAB}$	$(a-1)(b-1)$	$\overbrace{\frac{SSAB}{(a-1)(b-1)}}^{MSAB}$	$\sigma^2 + n\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot})^2}{(a-1)(b-1)}$
Error	$\overbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij\cdot})^2}^{SSE}$	$ab(n-1)$	$\overbrace{\frac{SSE}{ab(n-1)}}^{MSE}$	σ^2
Total	$\overbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2}^{SSTO}$	$abn-1$		

Attention to the $E[MS]$: motivation to build the test!

F tests

We can conduct a number of F tests, according to which effects' nullity we want to test. We can be interested in testing for the interaction effect or for either of the main effects. As we did for single-case study, we can understand the motivation for constructing the right F tests just by

- stating the null and alternative hypotheses of interest, to identify the theoretical quantity of interest,
- looking at the ANOVA table (and in particular at the $E[MS]$), to understand which is the appropriate test statistic to use.

Test for interactions

Our null and alternative hypotheses are, in this case:

$$\begin{cases} H_0: & (\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot} = 0 \quad \text{for all } i,\, j \\ H_1: & (\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot} \neq 0 \quad \text{for some } i,\, j \end{cases},$$

If we go back to our ANOVA table we notice that

- $E[MSAB] = \sigma^2 + n\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij}-\mu_{i\cdot}-\mu_{\cdot j}+\mu_{\cdot\cdot})^2}{(a-1)(b-1)}$,
- $E[MSE] = \sigma^2$.

This makes the ratio $F_{AB} = MSAB/MSE$ being in average equal to 1 under H_0 and larger than 1 under H_1. Hence large values of F_{AB} indicate the existence of interaction. Under H_0, F_{AB} is distributed as $F_{(a-1)(b-1),(n-1)ab}$, and we reject H_0 when F_{AB}^{obs} takes large values, which means that our rejection region R is on the right tail of the distribution. For a fixed level of significance α, $R = \left\{F : F_{AB}^{obs} > F_{(1-\alpha;(a-1)(b-1),(n-1)ab)}\right\}$.

Test Factors A and B main effects

For factor A, our null and alternative hypotheses are, in this case:

$$\begin{cases} H_0: & \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{a\cdot} \\ H_1: & \text{not all } \mu_{i\cdot} \text{ are equal} \end{cases},$$

or equivalently

$$\begin{cases} H_0: & \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \\ H_1: & \text{not all } \alpha_i \text{ equal } 0 \end{cases},$$

If we go back to our ANOVA table we notice that

- $E[MSA] = \sigma^2 + bn\frac{\sum_{i=1}^{a}(\mu_{i\cdot}-\mu_{\cdot\cdot})^2}{a-1}$,
- $E[MSE] = \sigma^2$.

This makes the ratio $F_A = MSA/MSE$ being in average equal to 1 under H_0 and larger than 1 under H_1. Hence large values of F_A indicate the existence of the main effect of Factor A. Under H_0, F_A is distributed as $F_{a-1,(n-1)ab}$, and we reject H_0 when F_A^{obs} takes large values, which means that our rejection region R is on the right tail of the distribution. For a fixed level of significance α, $R = \left\{F : F_A^{obs} > F_{(1-\alpha;a-1,(n-1)ab)}\right\}$. The same idea can be used to test for the Factor B main effect. We an use the ratio $F_B = MSB/MSE$, which is average equal to 1 under H_0 and larger than 1 under H_1. Hence large values of F_B indicate the existence of the main effect of Factor B. Under H_0, F_B is distributed as $F_{b-1,(n-1)ab}$, and, again, we reject H_0 when F_B^{obs} takes large values, This means that our rejection region R is, for a fixed level of significance α, $R = \left\{F : F_B^{obs} > F_{(1-\alpha;b-1,(n-1)ab)}\right\}$.

1.4 Analysis of factor levels effects with no interaction

After the model validation

We already discussed how an ANOVA problem should be solved by steps (in single-factor studies). In two-factor studies we have to:

- check the assumptions of the model (through the analysis of residuals, for instance),
- perform the F tests to check which effects (main factor and/or interaction) are playing an important role,
- analyze the single effects (estimation and testing), taking care of the multiplicity issue.

We can distinguish between the analysis of factor effects when factors do or do not interact.

Estimation of factor level mean

Unbiased point estimators are given by (the related variances are also reported):

Parameter	Estimator	Estimator's variance	Estimate of the variance
$\mu_{i\cdot}$	$\hat{\mu}_{i\cdot} = \bar{Y}_{i\cdot\cdot}$	$V\left[\bar{Y}_{i\cdot\cdot}\right] = \frac{\sigma^2}{bn}$	$s^2_{\bar{Y}_{i\cdot\cdot}} = \frac{MSE}{bn}$
$\mu_{\cdot j}$	$\hat{\mu}_{\cdot j} = \bar{Y}_{\cdot j\cdot}$	$V\left[\bar{Y}_{\cdot j\cdot}\right] = \frac{\sigma^2}{an}$	$s^2_{\bar{Y}_{\cdot j\cdot}} = \frac{MSE}{an}$

Confidence intervals can be built, as usual, using the t distribution as, at a significance level α:

$$\bar{y}_{i\cdot\cdot} \pm t_{(1-\alpha/2;(n-1)ab)} s_{\bar{Y}_{i\cdot\cdot}},$$
$$\bar{y}_{\cdot j\cdot} \pm t_{(1-\alpha/2;(n-1)ab)} s_{\bar{Y}_{\cdot j\cdot}}$$

Estimation of contrast (and linear combination) of factor level means

A contrast of factor level means $\mu_{i\cdot}$ is $L = \sum_{i=1}^{a} c_i \mu_{i\cdot}$, with $\sum_{i=1}^{a} c_i = 0$ (only for contrasts). Thus:

- unbiased point estimator is given by $\hat{L} = \sum_{i=1}^{a} c_i \bar{Y}_{i\cdot\cdot}$,
- the variance of $\hat{L}$ is $V\left[\hat{L}\right] = \sum_{i=1}^{a} c_i^2 V\left[\bar{Y}_{i\cdot\cdot}\right] = \frac{\sigma^2}{bn} \sum_{i=1}^{a} c_i^2$, which is estimated by $s^2_{\hat{L}} = \frac{MSE}{bn} \sum_{i=1}^{a} c_i^2$.

Same idea applies for contrasts of factor level means $\mu_{\cdot j}$, for which we use the estimator $\hat{L} = \sum_{j=1}^{b} c_j \bar{Y}_{\cdot j\cdot}$, and its variance estimate $s^2_{\hat{L}} = \frac{MSE}{an} \sum_{j=1}^{b} c_j^2$. Confidence intervals can be built as, at a significance level α:

$$\hat{l} \pm t_{(1-\alpha/2;(n-1)ab)} s_{\hat{L}}$$

Multiple pairwise comparisons or contrasts of factor level means

When multiple comparisons need to be performed, procedures that take care of the multiplicity issue need to be considered. As for single-case studies, different multiple comparisons procedures might be more appropriate according to the situations:

- *Tukey* procedure,
- *Bonferroni* procedure,
- *Sheffé* procedure.

Moreover, because two factors are present in the study, we might be interested to the *combined Factor A and Factor B family*, e.g. the family of pairwise comparisons involving both factors means.

Tukey procedure

To refresh our minds, this procedure can be applied when the family of interest is the set of *all pairwise comparisons* of factor levels means. Hence the global hypotheses are, for instance when factor A is of interest:

$$\begin{cases} H_{0\,glob}: & \bigcap_{m=1}^{M} H_{0\,m} \\ H_{1\,glob}: & \bigcup_{m=1}^{M} H_{1\,m} \end{cases} =$$

$$\begin{cases} H_{0\,glob}: & \bigcap_{i,i'=1}^{r} D = \mu_{i\cdot} - \mu_{i'\cdot} = 0 \\ H_{1\,glob}: & \bigcup_{i,i'=1}^{r} D = \mu_{i\cdot} - \mu_{i'\cdot} \neq 0 \end{cases}, \; i \neq i'.$$

We can build a CI for any pairwise comparison as:

$$\hat{d} \pm \frac{1}{\sqrt{2}} q_{(1-\alpha;a,(n-1)ab)} s_{\hat{D}}, \quad \text{for factor } A,$$
$$\hat{d} \pm \frac{1}{\sqrt{2}} q_{(1-\alpha;b,(n-1)ab)} s_{\hat{D}}, \quad \text{for factor } B.$$

where $q_{(1-\alpha;a,(n-1)ab)}$ and $q_{(1-\alpha;b,(n-1)ab)}$ are the $(1-\alpha)\%$ quantiles of a studentized range distribution, $\hat{d} = \bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}$ and $s^2_{\hat{D}} = 2MSE/bn$ for factor A, and $\hat{d} = \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot j'\cdot}$ and $s^2_{\hat{D}} = 2MSE/an$ for factor B. We can also perform a test for the null hypothesis $H_0 : D = 0$, and reject it if (for a two-sided alternative)

$$\left| \frac{\sqrt{2}\hat{d}}{s_{\hat{D}}} \right| > q_{(1-\alpha;a,(n-1)ab)}, \quad \text{for factor } A,$$

$$\left| \frac{\sqrt{2}\hat{d}}{s_{\hat{D}}} \right| > q_{(1-\alpha;b,(n-1)ab)}, \quad \text{for factor } B.$$

Sheffé procedure

This procedure can be applied when the family of interest is the set of *all possible contrasts* of factor levels means. Hence the global hypotheses are (for factors A or B)

$$\begin{cases} H_{0\,glob} : & \bigcap_{m=1}^{M} H_{0\,m} \\ H_{1\,glob} : & \bigcup_{m=1}^{M} H_{1\,m} \end{cases} =$$

$$\begin{cases} H_{0\,glob} : & \bigcap_{m=1}^{M_L} L_m = 0 \\ H_{1\,glob} : & \bigcup_{m=1}^{M_L} L_m \neq 0 \end{cases} .$$

We can build a CI for any contrast as:

$$\hat{l} \pm \sqrt{(a-1)\, F_{(1-\alpha;a-1,(n-1)ab)}} s_{\hat{L}}, \quad \text{for factor } A,$$

$$\hat{l} \pm \sqrt{(b-1)\, F_{(1-\alpha;b-1,(n-1)ab)}} s_{\hat{L}}, \quad \text{for factor } B,$$

where $s_{\hat{L}}$ are defined as we just saw, depending on which of the two factors A or B is considered. We can also perform a test for the null hypothesis $H_0 : L = 0$, and reject it if (for a two-sided alternative)

$$\frac{\hat{l}^2}{(a-1)\, s^2_{\hat{L}}} > F_{(1-\alpha;a-1,(n-1)ab)}, \quad \text{for factor } A,$$

$$\frac{\hat{l}^2}{(b-1)\, s^2_{\hat{L}}} > F_{(1-\alpha;b-1,(n-1)ab)}, \quad \text{for factor } B.$$

Bonferroni procedure

This procedure can be applied when the family of interest is a set of M_e *pairwise comparisons*, *contrasts* and/or *linear combinations* of factor levels means, which is specified by the analyst in advance of the data analysis. Hence the global hypotheses are (for factors A or B)

$$\begin{cases} H_{0\,glob}: & \bigcap_{m=1}^{M} H_{0\,m} \\ H_{1\,glob}: & \bigcup_{m=1}^{M} H_{1\,m} \end{cases} =$$

$$\begin{cases} H_{0\,glob}: & \bigcap_{m=1}^{M_e} L_m = 0 \\ H_{1\,glob}: & \bigcup_{m=1}^{M_e} L_m \neq 0 \end{cases} .$$

We can build a CI for any contrast as (for factors A or B):

$$\hat{l} \pm t_{(1-\alpha/(2M_e);(n-1)ab)} s_{\hat{L}},$$

Notice that the only difference with the non corrected CIs for L is the confidence level in $t_{(1-\alpha/(2M_e);(n-1)ab)}$, which is increased from $1-\alpha/2$ to $1-\alpha/(2M_e)$. We can also perform a test for the null hypothesis $H_0 : L = 0$, and reject it if (for a two-sided alternative)

$$\left| \frac{\hat{l}}{s_{\hat{L}}} \right| > t_{(1-\alpha/(2M_e);(n-1)ab)},$$

Combined factor A and factor B family

When both factor A and B have an effect, it might be of interest to control the inferential errors of the whole family of comparisons or contrasts, related to both factor means. We can do this by adopting different strategies:

- we can use directly the Bonferroni procedure, with M_e representing then the total number of statements in the joint set,
- we can use the Bonferroni method in conjunction with the Tukey or Sheffé method (according to which is the most appropriate between these two for the single factor problem). In practice the Bonferroni inequality will assure an upper bound for the actual global significance level,
- when contrasts of factor level means are of interest, the Sheffé procedure can be used with a slightly modified formula:
  - for the confidence intervals

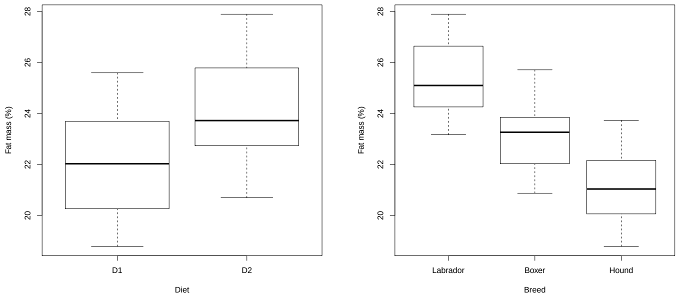
$$\hat{l} \pm \sqrt{(a+b-2)\, F_{(1-\alpha;a+b-2,(n-1)ab)}} s_{\hat{L}},$$

  - and for tests

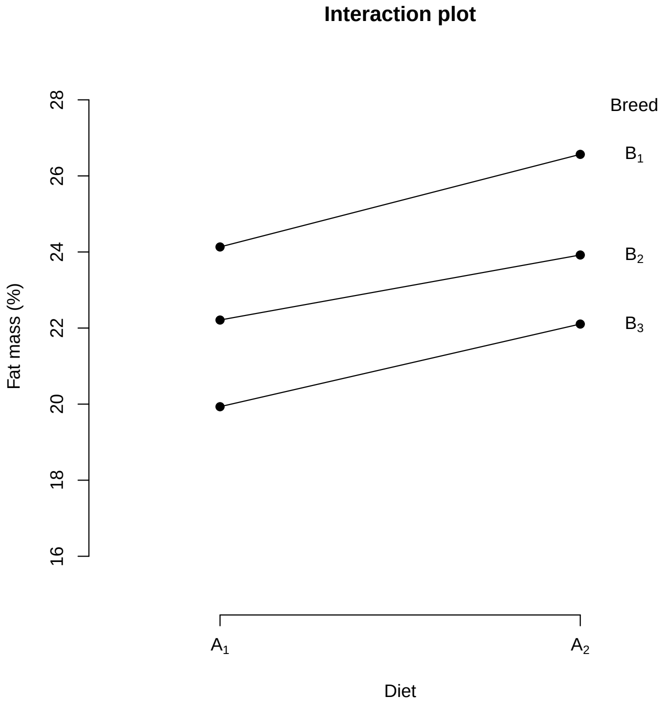
$$\left| \frac{\hat{l}^2}{(a+b-2)\, s_{\hat{L}}^2} \right| > F_{(1-\alpha;a+b-2,(n-1)ab)}.$$

Analyzing our example with R

We can consider again the example 2 about our study on dogs (main effects of both factors A and B and no interaction) We can simulate the data according to the cell means model, and obtain data looking like (boxplots of fat mass):



The interaction plot obtained from the estimates of the treatment means $\hat{\mu}_{ij}$ suggests the absence of interaction:



We can fit the initial two-way ANOVA model, with interaction:

```
> ## Estimating the models and performing the F tests:
> # define the contrasts
> contrasts(data_dogs$factorA) = contr.sum
> contrasts(data_dogs$factorB) = contr.sum
> # fit the initial model
> mod1_dogs_ex2 = lm(fat~factorA*factorB,data=data_dogs)

> summary(mod1_dogs_ex2)

Call:
lm(formula = fat ~ factorA * factorB, data = data_dogs)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84778 -0.74117  0.00157  0.45688  1.79029

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.14488    0.11991 193.012  < 2e-16 ***
factorA1          -1.05280    0.11991  -8.780 5.55e-12 ***
factorB1           2.20449    0.16958  12.999  < 2e-16 ***
factorB2          -0.07873    0.16958  -0.464    0.644
factorA1:factorB1 -0.16436    0.16958  -0.969    0.337
factorA1:factorB2  0.19780    0.16958   1.166    0.249
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9288 on 54 degrees of freedom
Multiple R-squared: 0.8458,     Adjusted R-squared: 0.8315
F-statistic: 59.24 on 5 and 54 DF,  p-value: < 2.2e-16
```

We can test the main effects and interaction, by performing the F tests:

```
> # perform the anova test
> anova(mod1_dogs_ex2)  # interaction effect is not significant
Analysis of Variance Table

Response: fat
                Df  Sum Sq Mean Sq  F value    Pr(>F)
factorA          1  66.504  66.504  77.0824 5.555e-12 ***
factorB          2 187.697  93.848 108.7770 < 2.2e-16 ***
factorA:factorB  2   1.345   0.673   0.7796    0.4637
Residuals       54  46.589   0.863
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can now test for all pairwise comparisons for factor B, for instance, using Tukey procedure. Ignoring factor A, we would have obtained:

```
> ## Performing multiple comparisons
> library(multcomp)
> # Tukey for factor B
> mod2_dogs_ex2_factorB = lm(fat~factorB,data=data_dogs)    # ignoring factor A
> tukey_factorB = glht(mod2_dogs_ex2_factorB, linfct = mcp(factorB = "Tukey"))
> tukey_ci_factorB = confint(tukey_factorB)
> tukey_ci_factorB

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = fat ~ factorB, data = data_dogs)

Quantile = 2.4066
95% family-wise confidence level
```

```
Linear Hypotheses:
                      Estimate lwr     upr
Boxer - Labrador == 0 -2.2832  -3.3615 -1.2049
Hound - Labrador == 0 -4.3303  -5.4086 -3.2519
Hound - Boxer == 0    -2.0470  -3.1254 -0.9687
```

But if we include factor A in the model estimation, we end up with tighter confidence intervals:

```
> tukey_factorB = glht(mod1_dogs_ex2, linfct = mcp(factorB = "Tukey")) # including factor A
+                                                                      # and interaction
Warning message:
In mcp2matrix(model, linfct = linfct) :
  covariate interactions found -- default contrast might be inappropriate
> tukey_ci_factorB = confint(tukey_factorB)
> tukey_ci_factorB

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = fat ~ factorA * factorB, data = data_dogs)

Quantile = 2.4103
95% family-wise confidence level


Linear Hypotheses:
                      Estimate lwr     upr
Boxer - Labrador == 0 -2.2832  -2.9912 -1.5752
Hound - Labrador == 0 -4.3303  -5.0382 -3.6223
Hound - Boxer == 0    -2.0470  -2.7550 -1.3391
```

If we drop the interaction term and retest for the main effects, we get slightly different results from the model with interaction. We are pooling the sums of squares: we need to be careful (see Kutner et al. (2005), Chapter 19.10.).

```
> mod2_dogs_ex2 = lm(fat~factorA+factorB,data=data_dogs)
>
> anova(mod2_dogs_ex2)  # both main effects are significant
Analysis of Variance Table

Response: fat
          Df  Sum Sq Mean Sq F value    Pr(>F)
factorA    1  66.504  66.504  77.694 3.603e-12 ***
factorB    2 187.697  93.848 109.640 < 2.2e-16 ***
Residuals 56  47.934   0.856
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1.5 Analysis of factor levels effects in presence of interaction

After the model validation

As we already mentioned, two different analysis must be distinguished when factors do or do not interact. The idea is that, when there is an interaction effect, it does not make sense anymore to compare the factor level means, since differences depend on the level of the other factor. Hence in this case, the analysis of factor effects generally must be based on the treatment means μ_{ij} (and no longer on $\mu_{i\cdot}$ and $\mu_{\cdot j}$). Alternatively one could be interested in comparing the levels of one factor across the levels of the other factors (also referred to a comparison of *simple effects*).

Multiple pairwise comparisons or contrasts of treatments means

Also in this case, when multiple comparisons need to be performed, procedures that take care of the multiplicity issue need to be considered. We know already that different multiple comparisons procedures might be more appropriate according to the situations, as *Tukey*, *Bonferroni* and *Sheffé* procedure. The choice of the procedure to be chosen should follow the same rules we already discussed.

Tukey procedure

This procedure can be applied when the family of interest is the set of *all pairwise comparisons* of treatment means, $D = \mu_{ij} - \mu_{i'j'}$. We can build a CI for any pairwise comparison as:

$$\hat{d} \pm \frac{1}{\sqrt{2}} q_{(1-\alpha;ab,(n-1)ab)} s_{\hat{D}},$$

where $s^2_{\hat{D}} = 2MSE/n$ and $\hat{d} = \bar{y}_{ij\cdot} - \bar{y}_{j'j'\cdot}$. We can also perform a test for the null hypothesis $H_0 : D = 0$, and reject it if (for a two-sided alternative)

$$\left| \frac{\sqrt{2}\hat{d}}{s_{\hat{D}}} \right| > q_{(1-\alpha;ab,(n-1)ab)}.$$

Sheffé procedure

This procedure can be applied when the family of interest is the set of *all possible contrasts* of treatment means, $L = \sum_{i=1}^{a} \sum_{j=1}^{b} c_{ij}\mu_{ij}$. We can build a CI for any contrast as:

$$\hat{l} \pm \sqrt{(ab-1)\, F_{(1-\alpha;ab-1,(n-1)ab)}} s_{\hat{L}},$$

where $s^2_{\hat{L}} = MSE/n \sum_{i=1}^{a} \sum_{j=1}^{b} c^2_{ij}$ and $\hat{l} = \sum_{i=1}^{a} \sum_{j=1}^{b} c_{ij}\bar{y}_{ij\cdot}$. We can also perform a test for the null hypothesis $H_0 : L = 0$, and reject it if (for a two-sided alternative)

$$\frac{\hat{l}^2}{(ab-1)\, s^2_{\hat{L}}} > F_{(1-\alpha;ab-1,(n-1)ab)}.$$

Bonferroni procedure

This procedure can be applied when the family of interest is a set of M_e *pairwise comparisons*, *contrasts* and/or *linear combinations* of factor levels means, which is specified by the analyst in advance of the data analysis. We can build a CI for any contrast as:

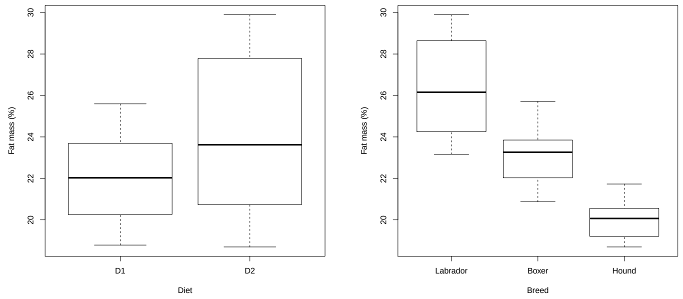
$$\hat{l} \pm t_{(1-\alpha/(2M_e);(n-1)ab)} s_{\hat{L}},$$

We can also perform a test for the null hypothesis $H_0 : L = 0$, and reject it if (for a two-sided alternative)

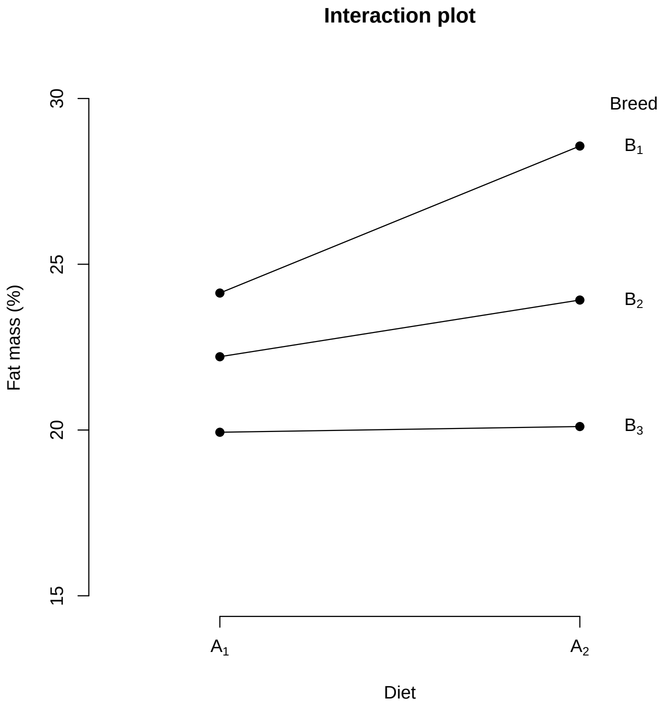
$$\left| \frac{\hat{l}}{s_{\hat{L}}} \right| > t_{(1-\alpha/(2M_e);(n-1)ab)},$$

Analyzing our example with R

We can consider again the example 3 about our study on dogs (main effects of both factors A and B and no interaction) We can simulate the data according to the cell means model, and obtain data looking like (boxplots of fat mass):



The interaction plot obtained from the estimates of the treatment means $\hat{\mu}_{ij}$ suggests the presence of interaction:



We can fit the initial two-way ANOVA model, with interaction:

```
> ## Estimating the models and performing the F tests:
> # define the contrasts
> contrasts(data_dogs$factorA) = contr.sum
> contrasts(data_dogs$factorB) = contr.sum
> # fit the initial model
> mod1_dogs_ex3 = lm(fat~factorA*factorB,data=data_dogs)

> summary(mod1_dogs_ex3)

Call:
lm(formula = fat ~ factorA * factorB, data = data_dogs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.84778 -0.74117  0.00157  0.45688  1.79029 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       23.14488    0.11991 193.012  < 2e-16 ***
factorA1          -1.05280    0.11991  -8.780 5.55e-12 ***
factorB1           3.20449    0.16958  18.896  < 2e-16 ***
factorB2          -0.07873    0.16958  -0.464    0.644    
factorA1:factorB1 -1.16436    0.16958  -6.866 6.80e-09 ***
factorA1:factorB2  0.19780    0.16958   1.166    0.249    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9288 on 54 degrees of freedom
Multiple R-squared: 0.9169,     Adjusted R-squared: 0.9092 
F-statistic: 119.2 on 5 and 54 DF,  p-value: < 2.2e-16
```

We can test the main effects and interaction, by performing the F tests:

```
> # perform the anova tests
> anova(mod1_dogs_ex3)  # interaction effect is now significant
Analysis of Variance Table

Response: fat
                Df Sum Sq Mean Sq F value    Pr(>F)    
factorA          1  66.50  66.504  77.082 5.555e-12 ***
factorB          2 400.91 200.454 232.340 < 2.2e-16 ***
factorA:factorB  2  46.58  23.291  26.996 7.466e-09 ***
Residuals       54  46.59   0.863                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can now test for all pairwise comparisons of treatment levels, for instance, using Tukey procedure (we don’t refer anymore to a specific factor):

```
> ## Performing multiple comparisons
> # Tukey all treatments
> data_dogs$treatments = data_dogs$factorA:data_dogs$factorB
> mod1_dogs_ex3_treatments = lm(fat~treatments,data=data_dogs)
> tukey_treatments = glht(mod1_dogs_ex3_treatments, linfct = mcp(treatments = "Tukey"))

> tukey_ci_treatments = confint(tukey_treatments)
> tukey_ci_treatments

         Simultaneous Confidence Intervals
```

```
Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = fat ~ treatments, data = data_dogs)

Quantile = 2.9536
95% family-wise confidence level


Linear Hypotheses:
                               Estimate lwr     upr
D1:Boxer - D1:Labrador == 0    -1.9211  -3.1485 -0.6936
D1:Hound - D1:Labrador == 0    -4.1993  -5.4268 -2.9719
D2:Labrador - D1:Labrador == 0  4.4343   3.2069  5.6617
D2:Boxer - D1:Labrador == 0    -0.2111  -1.4385  1.0164
D2:Hound - D1:Labrador == 0    -4.0268  -5.2543 -2.7994
D1:Hound - D1:Boxer == 0       -2.2783  -3.5057 -1.0509

D2:Labrador - D1:Boxer == 0     6.3554   5.1280  7.5828
D2:Boxer - D1:Boxer == 0        1.7100   0.4826  2.9374
D2:Hound - D1:Boxer == 0       -2.1058  -3.3332 -0.8784
D2:Labrador - D1:Hound == 0     8.6337   7.4062  9.8611
D2:Boxer - D1:Hound == 0        3.9883   2.7609  5.2157
D2:Hound - D1:Hound == 0        0.1725  -1.0549  1.3999
D2:Boxer - D2:Labrador == 0    -4.6454  -5.8728 -3.4180
D2:Hound - D2:Labrador == 0    -8.4612  -9.6886 -7.2338
D2:Hound - D2:Boxer == 0       -3.8158  -5.0432 -2.5884
```

ANALYSIS OF VARIANCE

Master of Statistics

Lesson 7

Dr. Francesca Solmi

Contents

1 Two-way ANOVA (Unequal sample size)

1.1 Two-way unbalanced ANOVA design

Two-way unbalanced ANOVA design
Kutner et al. (2005), Chapter 23.

So far, we focused on single-factor studies and two-factors studies with equal sample sizes. Now we go a step further into the study of two-factors studies with unequal sample sizes. Again, the simultaneous effect of the two factors is of interest. There are many reasons for ending up with unequal treatment sample sizes:

- in observational studies: because there might be not equal amounts of units available for the different factors' combinations.
- in experimental studies: patients can get sick at the moment of recording, the recording process might give problems and some information gets lost, ethical reasons might request the exclusions of some unit from the analysis, etc..
- in both observational and experimental studies: more units might be used for specific treatments which are cheaper to administrate, or to allow a more precise estimation of specific treatments effects.

We will use the same notation of lessons 5 and 6, with the only difference that now the sample size for the combination of the ith level of factor A and the jth level of factor B will be denoted with n_{ij}. Thus the total number of cases for the ith level of factor A is denoted by

$$n_{i\cdot} = \sum_{j=1}^{b} n_{ij},$$

the total number of cases for the jth level of factor B is denoted by

$$n_{\cdot j} = \sum_{i=1}^{a} n_{ij}$$

and the total number of cases for the entire study is denoted by

$$n_T = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}.$$

In this setting the estimates for the treatment mean when treatments A is at level i and treatment B is at level j are given by:

$$\hat{\mu}_{ij} = \bar{y}_{ij\cdot} = \frac{\sum_{k=1}^{n_{ij}} y_{ijk}}{n_{ij}}$$

Notice that, as in lessons 5 and 6, we denote the single observation as y_{ijk}, where the indexes $i = 1, \ldots, a$ and $j = 1, \ldots, b$ refer to the factor levels, and the index $k = 1, \ldots, n_{ij}$ denote the single observation for the specific treatment ij. Adopting this notation is in this case useful because the solution to the problem will be found in the linear regression approach.

Use of regression approach

The solution to problem of testing factor effects when the sample sizes are not equal cannot be simply extended from the one in presence of equal sample sizes. The problem is that the same variance decomposition no longer holds in this case. Hence we need to adopt another approach in order to solve the problem. Reminding that the two-way ANOVA model is nothing else than a particular regression model where the covariates are categorical, then an easy solution to the problem is represented by the classical linear regression approach.

1.2 Use of regression approach

Factor effect model

Let us consider the factor effect model:

$$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \ \ i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, n_{ij},$$

where

- $\mu_{..}$ is a parameter,
- α_i and β_j are the main factor effects parameters, subject to restrictions $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$,
- $(\alpha\beta)_{ij}$ are the interaction parameters, subject to the restriction

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 \quad j = 1, \dots, b, \sum_{j=1}^{b} (\alpha\beta)_{ij} \qquad\qquad = 0 \quad i = 1, \dots, a,$$

- ϵ_{ijk} are independent realizations from $N(0, \sigma^2)$.

We remind that the model has the properties:

- $E[Y_{ijk}] = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$,
- $V[Y_{ijk}] = \sigma^2$, *constant*,
- Y_{ijk} are *normally distributed*,
- y_{ijk} are *independent* realizations from the corresponding distributions.

The constraint on the α_i and β_j parameters

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$$

imply that the parameters we need to estimate in the model for the main factor effects are $(a-1)+(b-1)$ of the α_i and β_j, since two of them can be expressed in terms of the others. For instance, we can drop the parameters

$$\alpha_a = -\alpha_1 - \alpha_2 - \cdots - \alpha_{a-1} \text{ and } \beta_b = -\beta_1 - \beta_2 - \cdots - \beta_{b-1}.$$

So, as in lesson 4, we can use, for α_i and β_j, respectively $(a-1)$ and $(b-1)$ variables that can take values 1, -1 and 0. For the interaction parameters the constraints

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 \quad j = 1, \dots, b, \sum_{j=1}^{b} (\alpha\beta)_{ij} \qquad\qquad = 0 \quad i = 1, \dots, a,$$

imply that the parameters we need to estimate in the model for the interaction effects are $(a-1)(b-1)$ of the $(\alpha\beta)_{ij}$, since some of them can be expressed in terms of the others. For instance, for each $i = 1, \dots, a$ and $j = 1, \dots, b$, we can drop the parameters

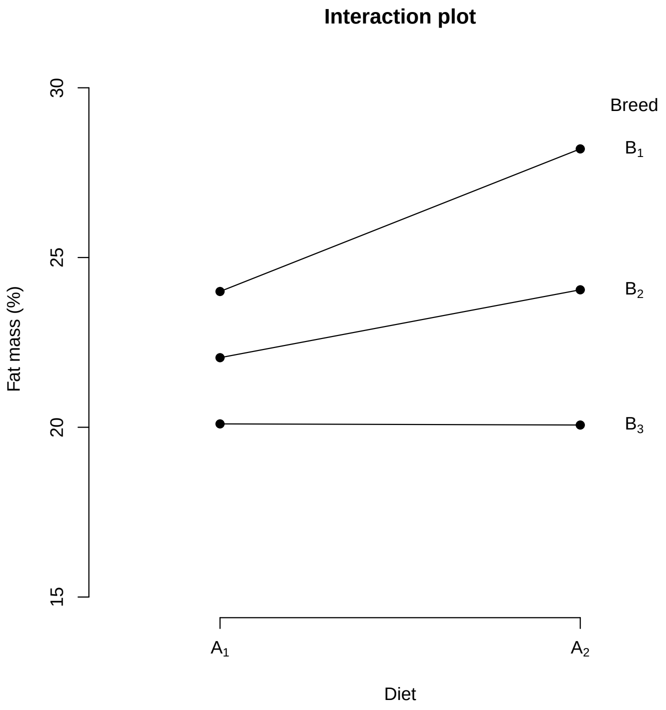
$$\begin{aligned}(\alpha\beta)_{ib} &= -(\alpha\beta)_{i1} - (\alpha\beta)_{i2} - \cdots - (\alpha\beta)_{i,b-1} \text{ and} \\ (\alpha\beta)_{aj} &= -(\alpha\beta)_{1j} - (\alpha\beta)_{2j} - \cdots - (\alpha\beta)_{a-1,j}.\end{aligned}$$

We will see that in the regression model we can associate to these terms the same indicator variables as for the main effects, by taking their cross products. Let us consider, as an example, the study on dogs from lessons 5 and 6. Thus we have $a = 2$ levels for factor A, $b = 3$ levels for factor B and the treatment sample sizes are now unequal:

	Factor B - Breed		
	$j = 1$	$j = 2$	$j = 3$
Factor A - Diet	**Labrador**	**Boxer**	**Hound**
$i = 1$ **D1**	22.3 (y_{111}) 25.6 (y_{112}) 24.1 (y_{113})	21.3 (y_{121}) 22.8 (y_{122})	19.1 (y_{131}) 21.1 (y_{132})
Mean	24 ($\bar{y}_{11\cdot}$)	22.05 ($\bar{y}_{12\cdot}$)	20.1 ($\bar{y}_{13\cdot}$)
$i = 2$ **D2**	28.2 (y_{211})	22.2 (y_{221}) 25.9 (y_{222})	19.5 (y_{231}) 20.4 (y_{232}) 20.3 (y_{233})
Mean	28.2 ($\bar{y}_{21\cdot}$)	24.05 ($\bar{y}_{22\cdot}$)	20.7 ($\bar{y}_{23\cdot}$)

The interaction plot obtained from the estimates of the treatment means $\hat{\mu}_{ij}$ appears as:

```
> ##### Dogs example
> ### Interaction plot
> fat_11 = c(22.3,25.6,24.1)
> fat_12 = c(21.3,22.8)
> fat_13 = c(19.1,21.1)
> fat_21 = c(28.2)
> fat_22 = c(22.2,25.9)
> fat_23 = c(19.5,20.4,20.3)
> fat = c(fat_11,fat_12,fat_13,fat_21,fat_22,fat_23)
> factorA = as.factor(c(rep("D1",7),rep("D2",6)))
> factorB = as.factor(c(rep("Labrador",3),rep("Boxer",2),rep("Hound",2),
+            rep("Labrador",1),rep("Boxer",2),rep("Hound",3)))
> data_dogs = data.frame(cbind("fat"=fat,
+                              "factorA"=factorA,
+                              "factorB"=factorB))
> data_dogs$factorA = factor(factorA,levels=c("D1","D2"))
> data_dogs$factorB = factor(factorB,levels=c("Labrador","Boxer","Hound"))
> str(data_dogs)
'data.frame':   13 obs. of  3 variables:
 $ fat    : num  22.3 25.6 24.1 21.3 22.8 19.1 21.1 28.2 22.2 25.9 ...
 $ factorA: Factor w/ 2 levels "D1","D2": 1 1 1 1 1 1 1 2 2 2 ...
 $ factorB: Factor w/ 3 levels "Labrador","Boxer",..: 1 1 1 2 2 3 3 1 2 2 ...

> hat_mu = array(c(mean(fat_11),mean(fat_21),
+                  mean(fat_12),mean(fat_22),
+                  mean(fat_13),mean(fat_23)),dim=c(2,3))
> plot(c(1,2),hat_mu[,1],main="Interaction plot",
+      xlab="Diet",ylab="Fat mass (%)",
+      type="l",axes=F,
+      ylim=c(15,max(hat_mu[2,])+2),
+      xlim=c(0.7,2.3))
> points(c(1,2),hat_mu[,1],pch=19)
> points(c(1,2),hat_mu[,2],type="l")
> points(c(1,2),hat_mu[,2],pch=19)
> points(c(1,2),hat_mu[,3],type="l")
> points(c(1,2),hat_mu[,3],pch=19)
> axis(1,at=c(1,2),labels=c(expression(A[1]),
+                           expression(A[2])))
> axis(2)
> text(2+0.15,hat_mu[2,1]+1.3,label="Breed")
> text(2+0.15,hat_mu[2,1],label=expression(B[1]))
> text(2+0.15,hat_mu[2,2],label=expression(B[2]))
> text(2+0.15,hat_mu[2,3],label=expression(B[3]))
```



We can now develop the corresponding linear model.

$$\overbrace{\begin{pmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{121} \\ y_{122} \\ y_{131} \\ y_{132} \\ y_{211} \\ y_{221} \\ y_{222} \\ y_{231} \\ y_{232} \\ y_{233} \end{pmatrix}}^{\boldsymbol{Y}} = \overbrace{\begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix}}^{\begin{array}{cccccc} \mathbf{1} & X_1^A & X_1^B & X_2^B & X_1^A X_1^B & X_1^A X_2^B \end{array}} \overbrace{\begin{pmatrix} \mu_{\cdot\cdot} \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \end{pmatrix}}^{\boldsymbol{\beta}} + \overbrace{\begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{113} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{131} \\ \epsilon_{132} \\ \epsilon_{211} \\ \epsilon_{221} \\ \epsilon_{222} \\ \epsilon_{231} \\ \epsilon_{232} \\ \epsilon_{233} \end{pmatrix}}^{\boldsymbol{\epsilon}}$$

$$= \begin{pmatrix} \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} + \epsilon_{111} \\ \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} + \epsilon_{112} \\ \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} + \epsilon_{113} \\ \mu_{\cdot\cdot} + \alpha_1 + \beta_2 + (\alpha\beta)_{12} + \epsilon_{121} \\ \mu_{\cdot\cdot} + \alpha_1 + \beta_2 + (\alpha\beta)_{12} + \epsilon_{122} \\ \mu_{\cdot\cdot} + \alpha_1 - \beta_1 - \beta_2 - (\alpha\beta)_{11} - (\alpha\beta)_{12} + \epsilon_{131} \\ \mu_{\cdot\cdot} + \alpha_1 - \beta_1 - \beta_2 - (\alpha\beta)_{11} - (\alpha\beta)_{12} + \epsilon_{132} \\ \mu_{\cdot\cdot} - \alpha_1 + \beta_1 - (\alpha\beta)_{11} + \epsilon_{211} \\ \mu_{\cdot\cdot} - \alpha_1 + \beta_2 - (\alpha\beta)_{12} + \epsilon_{221} \\ \mu_{\cdot\cdot} - \alpha_1 + \beta_2 - (\alpha\beta)_{12} + \epsilon_{222} \\ \mu_{\cdot\cdot} - \alpha_1 - \beta_1 - \beta_2 + (\alpha\beta)_{11} + (\alpha\beta)_{12} + \epsilon_{231} \\ \mu_{\cdot\cdot} - \alpha_1 - \beta_1 - \beta_2 + (\alpha\beta)_{11} + (\alpha\beta)_{12} + \epsilon_{232} \\ \mu_{\cdot\cdot} - \alpha_1 - \beta_1 - \beta_2 + (\alpha\beta)_{11} + (\alpha\beta)_{12} + \epsilon_{233} \end{pmatrix}$$

Thus, in general for a two-factor study with a levels for factor A and b levels for factor B, the multiple regression model can be written as

$$y_{ijk} = \mu_{\cdot\cdot} + \sum_{i=1}^{a-1} \alpha_i x_{ijk,i}^A + \sum_{j=1}^{b-1} \beta_j x_{ijk,j}^B + \sum_{i=1}^{a-1} \sum_{j=1}^{b-1} (\alpha\beta)_{ij} x_{ijk,i}^A x_{ijk,j}^B + \epsilon_{ijk},$$

where the covariates are defined as

$$x_{ijk,1}^A = \begin{cases} 1 & \text{if case from factor } A \text{ level } 1 \\ -1 & \text{if case from factor } A \text{ level } a \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$x_{ijk,a-1}^A = \begin{cases} 1 & \text{if case from factor } A \text{ level } a-1 \\ -1 & \text{if case from factor } A \text{ level } a \\ 0 & \text{otherwise} \end{cases} .$$

and

$$x_{ijk,1}^B = \begin{cases} 1 & \text{if case from factor } B \text{ level } 1 \\ -1 & \text{if case from factor } B \text{ level } b \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$x_{ijk,b-1}^B = \begin{cases} 1 & \text{if case from factor } B \text{ level } b-1 \\ -1 & \text{if case from factor } B \text{ level } b \\ 0 & \text{otherwise} \end{cases} .$$

Hence, in terms of regression model, $\mu_{\cdot\cdot}$ is the intercept term, and $\alpha_i, \beta_j, (\alpha\beta)_{ij}$, $i = 1, \ldots, a-1$ and $j = 1, \ldots, b-1$ are the regression parameters.

Cell means model

Let us consider the cell means model:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n_{ij},$$

where

- μ_{ij} are parameters,
- ϵ_{ijk} are independent realizations from $N(0, \sigma^2)$

Again, the model has the properties:

- $E[Y_{ijk}] = \mu_{ij}$,
- $V[Y_{ijk}] = \sigma^2$, *constant*,
- Y_{ijk} are *normally distributed*,
- y_{ijk} are *independent* realizations from the corresponding distributions.

The corresponding linear model for the dogs example is then

$$\begin{aligned}
\overbrace{\begin{pmatrix} y_{111}\\ y_{112}\\ y_{113}\\ y_{121}\\ y_{122}\\ y_{131}\\ y_{132}\\ y_{211}\\ y_{221}\\ y_{222}\\ y_{231}\\ y_{232}\\ y_{233} \end{pmatrix}}^{\boldsymbol{Y}}
&= \overbrace{\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 0 & 0\\
0 & 1 & 0 & 0 & 0 & 0\\
0 & 1 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 0 & 0\\
0 & 0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 0 & 0 & 1\\
0 & 0 & 0 & 0 & 0 & 1\\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}}^{\begin{array}{cccccc} X_{11} & X_{12} & X_{13} & X_{21} & X_{22} & X_{23} \end{array}}
\overbrace{\begin{pmatrix} \mu_{11}\\ \mu_{12}\\ \mu_{13}\\ \mu_{21}\\ \mu_{22}\\ \mu_{23} \end{pmatrix}}^{\boldsymbol{\beta}}
+ \overbrace{\begin{pmatrix} \epsilon_{111}\\ \epsilon_{112}\\ \epsilon_{113}\\ \epsilon_{121}\\ \epsilon_{122}\\ \epsilon_{131}\\ \epsilon_{132}\\ \epsilon_{211}\\ \epsilon_{221}\\ \epsilon_{222}\\ \epsilon_{231}\\ \epsilon_{232}\\ \epsilon_{233} \end{pmatrix}}^{\boldsymbol{\epsilon}} \\
&= \begin{pmatrix} \mu_{11} + \epsilon_{111}\\ \mu_{11} + \epsilon_{112}\\ \mu_{11} + \epsilon_{113}\\ \mu_{12} + \epsilon_{121}\\ \mu_{12} + \epsilon_{122}\\ \mu_{13} + \epsilon_{131}\\ \mu_{13} + \epsilon_{132}\\ \mu_{21} + \epsilon_{211}\\ \mu_{22} + \epsilon_{221}\\ \mu_{22} + \epsilon_{222}\\ \mu_{23} + \epsilon_{231}\\ \mu_{23} + \epsilon_{232}\\ \mu_{23} + \epsilon_{233} \end{pmatrix}
\end{aligned}$$

Thus, in general for a two-factor study with a levels for factor A and b levels for factor B, the multiple regression model can be written as

$$y_{ijk} = \sum_{i=1}^{a} \sum_{j=1}^{b} \mu_{ij} x_{ijk,ij} + \epsilon_{ijk},$$

where the covariates are defined as

$$x_{ijk,ij} = \begin{cases} 1 & \text{if case from factor } A \text{ level } i \text{ and factor } B \text{ level } j \\ 0 & \text{otherwise} \end{cases}$$

Hence, in terms of regression model, μ_{ij}, $i = 1, \ldots, a$ and $j = 1, \ldots, b$ are the regression parameters.

1.3 Testing for interaction and main effects

Test for interaction effects

We can easily test for interaction and main effects using the *factor effect model.* The null and alternative hypotheses to test for the interaction effect can be stated in this case as

$$\begin{cases} H_0: & (\alpha\beta)_{ij} = 0 \quad \text{for all } i,\, j \\ H_1: & (\alpha\beta)_{ij} \neq 0 \quad \text{for some } i,\, j \end{cases}.$$

In the dogs example these hypotheses become

$$\begin{cases} H_0: & (\alpha\beta)_{11} = (\alpha\beta)_{12} = 0 \\ H_1: & \text{not both } (\alpha\beta)_{11} \text{ and } (\alpha\beta)_{12} \text{ equal } 0 \end{cases}.$$

This corresponds to simply test whether two regression coefficients of our model equal 0. This problem can be solved by performing the corresponding F test to compare the full factor effects model with the reduced model

$$y_{ijk} = \mu_{..} + \alpha_1 x^A_{ijk,1} + \beta_1 x^B_{ijk,1} + \beta_2 x^B_{ijk,2} + \epsilon_{ijk},$$

where the interaction terms are missing. The test statistic is defined as

$$F^{regr} = \frac{SSE_r - SSE_f}{\underbrace{df_r - df_f}_{(n_T - p_r)-(n_T - p_f) = p_f - p_r}} \Big/ \frac{SSE_f}{\underbrace{df_f}_{n_T - p_f}},$$

where SSE_r, df_r, p_r and SSE_f, df_f, p_f are respectively the *error sums of squares*, *degrees of freedom*, *number of regression parameters* for the reduced and the full model. The observed value of F^{regr} needs to be compared with the null distribution $F_{((p_f - p_r, n_T - p_f))}$. Large values of F^{regr} are evidence in favour of H_1. We can use R to compute the sums of squares and perform the test. We should first set the right model parametrization (factor effects model), setting the right contrasts in R. Then a linear model can be estimated:

```
> ### Testing the interaction effect
> ## Computing SSE_r and SSE_f and performing the test
> # define the contrasts (factor effects model)
> contrasts(data_dogs$factorA) = contr.sum
> contrasts(data_dogs$factorB) = contr.sum

> # fit the full and reduced models
> mod_f = lm(fat~factorA*factorB,data=data_dogs)
> summary(mod_f)

Call:
lm(formula = fat ~ factorA * factorB, data = data_dogs)

Residuals:
   Min     1Q Median     3Q    Max 
 -1.85  -0.75   0.10   0.75   1.85 

Coefficients:
```

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.07778    0.44723  51.602 2.69e-10 ***
factorA1          -1.02778    0.44723  -2.298  0.05514 .
factorB1           3.02222    0.67280   4.492  0.00283 **
factorB2          -0.02778    0.62410  -0.045  0.96574
factorA1:factorB1 -1.07222    0.67280  -1.594  0.15504
factorA1:factorB2  0.02778    0.62410   0.045  0.96574
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.508 on 7 degrees of freedom
Multiple R-squared: 0.8224,     Adjusted R-squared: 0.6956
F-statistic: 6.485 on 5 and 7 DF,  p-value: 0.01465

> sum(mod_f$residuals^2)
[1] 15.91667

> mod_r = lm(fat~factorA+factorB,data=data_dogs)
> summary(mod_r)

Call:
lm(formula = fat ~ factorA + factorB, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3169 -0.8839 -0.3729  0.9830  2.0593

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.8133     0.4607  49.517  2.8e-12 ***
factorA1     -0.8661     0.4783  -1.811  0.10361
factorB1      2.6698     0.6867   3.888  0.00369 **
factorB2      0.2367     0.6612   0.358  0.72859
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.643 on 9 degrees of freedom
Multiple R-squared: 0.729,      Adjusted R-squared: 0.6386
F-statistic: 8.069 on 3 and 9 DF,  p-value: 0.006409

> sum(mod_r$residuals^2)
[1] 24.29644

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_r = length(coef(mod_r))
> n_T
[1] 13
> p_f
[1] 6
> p_r
[1] 4
> f_regr = ((sum(mod_r$residuals^2)-sum(mod_f$residuals^2))/(p_f-p_r))/
+          (sum(mod_f$residuals^2)/(n_T-p_f))
> pvalue_interaction = 1-pf(f_regr,p_f-p_r,n_T-p_f)
> pvalue_interaction
[1] 0.2275533
```

Hence we conclude in this case that there is not enough evidence to say that there is significant interaction between the two factors.

Test for main effects

We can also test whether factor A and B main effects are present. The null and alternative hypotheses to test for the main effects can be stated as

$$\begin{cases} H_0: & \alpha_1 = \cdots = \alpha_{a-1} = 0 \\ H_1: & \alpha_i \neq 0 \quad \text{for some } i = 1, \ldots, a-1 \end{cases} .$$

and

$$\begin{cases} H_0: & \beta_1 = \cdots = \beta_{b-1} = 0 \\ H_1: & \beta_j \neq 0 \quad \text{for some } j = 1, \ldots, b-1 \end{cases} .$$

In the dogs example these hypotheses become

$$\begin{cases} H_0: & \alpha_1 = 0 \\ H_1: & \alpha_1 \neq 0 \end{cases} .$$

and

$$\begin{cases} H_0: & \beta_1 = \beta_2 = 0 \\ H_1: & \text{not both } \beta_1 \text{ and } \beta_2 \text{ equal } 0 \end{cases} .$$

This corresponds again to simply test whether some specific regression coefficients of our model equal 0. This problem can be solved by performing the corresponding F test to compare the full cell mean model with the reduced models

$$\begin{aligned} y_{ijk} =& \mu_{\cdot\cdot} + \beta_1 x^B_{ijk,1} + \beta_2 x^B_{ijk,2} \\ &+ (\alpha\beta)_{11} x^A_{ijk,1} x^B_{ijk,1} + (\alpha\beta)_{12} x^A_{ijk,1} x^B_{ijk,2} + \epsilon_{ijk} \end{aligned}$$

or

$$y_{ijk} = \mu_{\cdot\cdot} + \alpha_1 x^A_{ijk,1} + (\alpha\beta)_{11} x^A_{ijk,1} x^B_{ijk,1} + (\alpha\beta)_{12} x^A_{ijk,1} x^B_{ijk,2} + \epsilon_{ijk}$$

where respectively the main effects of factor A and B are missing. The test statistic is defined as before

$$F^{regr} = \frac{SSE_r - SSE_f}{\underbrace{df_r - df_f}_{(n_T - p_r) - (n - T - p_f) = p_f - p_r}} \Big/ \frac{SSE_f}{\underbrace{df_f}_{n_T - p_f}},$$

Performing the analysis in R, we obtain

```
> ### Testing the main effects
> ## Computing SSE_r and SSE_f and performing the test
> # fit the reduced model without factor A
> X0 <- model.matrix(mod_f)
> X1 <- X0[,!colnames(X0) %in% "factorA1"]  # we need to modify the
>                                           # matrix of regressors by hand
>                                           # (R tends to reparametrize the
>                                           #  model otherwise)
>
> mod_rA = lm(fat~0+X1,data=data_dogs)
> summary(mod_rA)

Call:
lm(formula = fat ~ 0 + X1, data = data_dogs)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.34912 -0.87632  0.02632  0.95088  2.82368 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
X1(Intercept)       22.91550    0.54717  41.880 1.16e-10 ***
X1factorB1           2.53538    0.79121   3.204   0.0125 *  
X1factorB2           0.13450    0.76830   0.175   0.8654    
X1factorA1:factorB1 -0.80175    0.82076  -0.977   0.3573    
X1factorA1:factorB2 -0.02632    0.77272  -0.034   0.9737    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.868 on 8 degrees of freedom
Multiple R-squared: 0.9958,     Adjusted R-squared: 0.9932 
F-statistic: 381.4 on 5 and 8 DF,  p-value: 2.735e-09

> sum(mod_rA$residuals^2)
[1] 27.92544

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))   # the full model is the same as before
> p_rA = length(coef(mod_rA))
> n_T
[1] 13
> p_f
[1] 6
> p_rA
[1] 5
> f_regrA = ((sum(mod_rA$residuals^2)-sum(mod_f$residuals^2))/(p_f-p_rA))/
+           (sum(mod_f$residuals^2)/(n_T-p_f))
> pvalue_factorA = 1-pf(f_regrA,p_f-p_rA,n_T-p_f)
> pvalue_factorA
[1] 0.05514384
> 
> 
> 
> 
> # fit the reduced model without factor B
> X1 <- X0[,(!colnames(X0) %in% "factorB1") & (!colnames(X0) %in% "factorB2")]
>                                                                   # we need to modify the
>                                                                   # matrix of regressors by hand
>                                                                   # (R tends to reparametrize the
>                                                                   #  model otherwise)
> 
> mod_rB = lm(fat~0+X1,data=data_dogs)

> summary(mod_rB)
```

```
Call:
lm(formula = fat ~ 0 + X1, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0424 -1.7949 -0.3186  1.8407  5.2220 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
X1(Intercept)       22.61864    0.88869  25.452 1.07e-09 ***
X1factorA1          -0.27853    0.85600  -0.325    0.752    
X1factorA1:factorB1 -0.08079    1.27586  -0.063    0.951    
X1factorA1:factorB2 -0.72147    1.22855  -0.587    0.571    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.053 on 9 degrees of freedom
Multiple R-squared: 0.9875,     Adjusted R-squared: 0.9819 
F-statistic: 177.1 on 4 and 9 DF,  p-value: 1.512e-08

> sum(mod_rB$residuals^2)
[1] 83.8739

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))   # the full model is the same as before
> p_rB = length(coef(mod_rB))
> n_T
[1] 13
> p_f
[1] 6
> p_rB
[1] 4
> f_regrB = ((sum(mod_rB$residuals^2)-sum(mod_f$residuals^2))/(p_f-p_rB))/
+           (sum(mod_f$residuals^2)/(n_T-p_f))
> pvalue_factorB = 1-pf(f_regrB,p_f-p_rB,n_T-p_f)
> pvalue_factorB
[1] 0.002977066
```

Hence we conclude in this case that there is a significant main effect of factor B. Factor A main effect results to be not significant at a 95% confidence level.

ANALYSIS OF VARIANCE
Master of Statistics

Lesson 8

Dr. Francesca Solmi

Contents

1 Estimable Functions

1.1 The problem

Motivating example

Consider a one-factor study, where the factor has $r = 3$ levels. Moreover assume we observe the response of interest on a sample with unequal sample sizes, $n_1 = 3$, $n_2 = 2$ and $n_3 = 2$. Consider the factor effect model formulation

$$y_{ij} = \mu. + \tau_i + \epsilon ij$$

with $i = 1, 2, 3$, $j = 1, \ldots, n_i$. Assume we choose to represent the model with the regression approach as:

$$\overbrace{\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}}^{\boldsymbol{Y}} = \overset{\overbrace{\mathbf{1} \quad X_1 \quad X_2 \quad X_3}^{\boldsymbol{X}}}{\overbrace{\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}}} \overbrace{\begin{pmatrix} \mu. \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}}^{\boldsymbol{\beta}} + \overbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}}^{\boldsymbol{\epsilon}}$$

From the theory of linear models we know that a solution to estimate $\boldsymbol{\beta}$ is given by

$$\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^{'}\boldsymbol{Y} \qquad \Longrightarrow \qquad \hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}$$

This means that we need to invert $\boldsymbol{X}^{'}\boldsymbol{X}$ in order to solve the estimation problem. Because of the way we formulated $\boldsymbol{X}$, the squared matrix $\boldsymbol{X}^{'}\boldsymbol{X}$ is not invertible in this case:

- $\boldsymbol{X}$ has not full rank ($rank(\boldsymbol{X}) = 3$),
- $\boldsymbol{X}^{'}\boldsymbol{X}$ has not full rank either,
- $\boldsymbol{X}^{'}\boldsymbol{X}$ is not invertible.

Thus, to solve the estimation problem, we need to use a *generalized inverse* of $\boldsymbol{X}^{'}\boldsymbol{X}$, namely $\boldsymbol{G}$. A *generalized inverse* of $\boldsymbol{X}^{'}\boldsymbol{X}$ is a matrix $\boldsymbol{G}$ satisfying

$$\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{X}^{'}\boldsymbol{X} \qquad \text{and} \qquad \boldsymbol{G}\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{G} = \boldsymbol{G}.$$

Then the solution for the estimation of $\boldsymbol{\beta}$ is given by

$$\begin{aligned} (\boldsymbol{X}^{'}\boldsymbol{X})\boldsymbol{\beta} &= \boldsymbol{X}^{'}\boldsymbol{Y} \\ (\boldsymbol{X}^{'}\boldsymbol{X})\boldsymbol{G}(\boldsymbol{X}^{'}\boldsymbol{X})\boldsymbol{\beta} &= (\boldsymbol{X}^{'}\boldsymbol{X})\boldsymbol{G}\boldsymbol{X}^{'}\boldsymbol{Y} \\ (\boldsymbol{X}^{'}\boldsymbol{X})\boldsymbol{\beta} &= (\boldsymbol{X}^{'}\boldsymbol{X})\boldsymbol{G}\boldsymbol{X}^{'}\boldsymbol{Y} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{G}} &= \boldsymbol{G}\boldsymbol{X}^{'}\boldsymbol{Y} \end{aligned}$$

Notice that we have $\boldsymbol{G}$ in $\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}$. Let us see an example. The following two matrices are both generalized inverses of our $(\boldsymbol{X}^{'}\boldsymbol{X})$:

$$\boldsymbol{G}_1 = \frac{1}{6}\begin{pmatrix} 3 & -3 & -3 & 0 \\ -3 & 5 & 3 & 0 \\ -3 & 3 & 6 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\boldsymbol{G}_2 = \frac{1}{6}\begin{pmatrix} 2 & 0 & -2 & -2 \\ 0 & 0 & 0 & 0 \\ -2 & 0 & 5 & 2 \\ -2 & 0 & 2 & 5 \end{pmatrix}.$$

These give rise to two possible solutions for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_1} = \boldsymbol{G}_1\boldsymbol{X}^{'}\boldsymbol{Y} = (81, -16, -11, 0)^{'}$$
$$\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_2} = \boldsymbol{G}_2\boldsymbol{X}^{'}\boldsymbol{Y} = (73, 0, 5, 16)^{'}$$

Let us look at the expected values of $\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_1}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_2}$:

$$E\left[\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_1}\right] = \boldsymbol{G}_1\boldsymbol{X}^{'}E(\boldsymbol{Y}) = \boldsymbol{G}_1\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{\beta} = \begin{pmatrix} \mu. + \tau_3 \\ \tau_1 - \tau_3 \\ \tau_2 - \tau_3 \\ 0 \end{pmatrix}$$

$$E\left[\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_2}\right] = \boldsymbol{G}_2\boldsymbol{X}^{'}E(\boldsymbol{Y}) = \boldsymbol{G}_2\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{\beta} = \begin{pmatrix} \mu. + \tau_1 \\ 0 \\ -\tau_1 + \tau_2 \\ -\tau_1 + \tau_3 \end{pmatrix}.$$

This explains the two different solutions. Hence, no unique solution exists for $\boldsymbol{\beta}$ in this case. How to solve the problem then? An idea is that of looking for linear combinations of $\boldsymbol{\beta}$ (if they exist) that can be uniquely estimated, regardless of the generalized inverse $\boldsymbol{G}$ used. Thus we need to search for linear combinations of $\boldsymbol{\beta}$, say $\boldsymbol{q}'\boldsymbol{\beta}$, such that $\boldsymbol{q}^{'}\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}$ is invariant to $\boldsymbol{G}$. If such linear combinations exist, then we can look directly at them after the estimation process (so unique conclusions ca be drawn). The good news is that such linear combinations exist and they are called *estimable functions*. For instance, in our case some invariant linear combinations of $\boldsymbol{\beta}$ are:

Parameters	$\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_1}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_2}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_3}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_4}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_5}$
$\mu.$	0	82.25	79	89	5283
τ_1	73	-9.25	-6	-16	-5210
τ_2	78	-4.25	-1	-11	-5205
τ_3	89	6.75	10	0	-5194

Linear Functions

$\boldsymbol{q}^{'}\boldsymbol{\beta}$	$\boldsymbol{q}^{'}\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_1}$	$\boldsymbol{q}^{'}\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_2}$	$\boldsymbol{q}^{'}\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_3}$	$\boldsymbol{q}^{'}\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_4}$	$\boldsymbol{q}^{'}\hat{\boldsymbol{\beta}}_{G_5}$
$\tau_1 + \tau_2$	151	-13.50	-7	-27	-10415
$\frac{1}{3}(\tau_1 + \tau_2 + \tau_3)$	80	-2.75	-1	-9	-5203
$\tau_1 - \tau_2$	-5	-5	-5	-5	-5
$\mu. + \tau_1$	73	73	73	73	73
$\mu. + \tau_3$	89	89	89	89	89

The following theorems will help in obtaining estimable functions.

Theorem 1 A function $\boldsymbol{q}^{'}\boldsymbol{\beta}$ for which $\boldsymbol{q}^{'}\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}$ is invariant to $\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}$ (or $\boldsymbol{G}$) is an estimable function if, and only if, $\boldsymbol{q}^{'} = \boldsymbol{t}^{'}\boldsymbol{X}$, for some $\boldsymbol{t} \in R^{n_T}$, where $\boldsymbol{X}$ has n_T rows.

Theorem 2 If $\boldsymbol{q}^{'}\boldsymbol{\beta}$ is an estimable function, then for every generalized inverse $\boldsymbol{G}$, $E\left(\boldsymbol{q}^{'}\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}\right) = \boldsymbol{q}^{'}\boldsymbol{\beta}$.

Let us check if one of the invariant linear combinations we saw can be written as $\boldsymbol{q}^{'}\boldsymbol{\beta}$, with $\boldsymbol{q}^{'} = \boldsymbol{t}^{'}\boldsymbol{X}$ (Theorem 1). Let us take, for instance, $\tau_1 - \tau_2$. We can write:

$$\tau_1 - \tau_2 = \overbrace{(0\ 1\ \ -1\ 0)}^{\boldsymbol{q}^{'}}\overbrace{\begin{pmatrix} \mu. \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}}^{\boldsymbol{\beta}}$$

We need to find a $\boldsymbol{t}^{'}$ such that $\boldsymbol{q}^{'} = \boldsymbol{t}^{'}\boldsymbol{X}$. We can write:

$$\overbrace{(0\ 1\ \ -1\ 0)}^{\boldsymbol{q}^{'}} = \overbrace{(1\ \ -1\ 1\ 0\ \ -1\ \ -1\ 1)}^{\boldsymbol{t}^{'}}\overbrace{\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}}^{\boldsymbol{X}}.$$

1.2 Basic Estimable Functions

Basic Estimable Functions

Starting from the matrix $\boldsymbol{X}$ we can write any estimable function. To do this, we first need to define the *basic estimable* functions. The general form for an estimable function is

$$\boldsymbol{q}^{'}\boldsymbol{\beta} = \overbrace{(t_1 \; t_2 \; \dots \; t_n)}^{\boldsymbol{t}^{'}} \overbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_T1} & x_{n_T2} & \dots & x_{n_Tp} \end{pmatrix}}^{\boldsymbol{X}} \boldsymbol{\beta}.$$

The *basic estimable* functions are generated by $\boldsymbol{t}^{'}$ vectors having all 0 and only one 1 in position l, $l = 1, \dots, n_T$:

$$\boldsymbol{q}^{'} = \overbrace{(0 \; \dots \; 0 \; 1 \; 0 \; \dots \; 0)}^{\boldsymbol{t}^{'}} \overbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_T1} & x_{n_T2} & \dots & x_{n_Tp} \end{pmatrix}}^{\boldsymbol{X}} = \overbrace{(x_{l1} \; x_{l2} \; \dots \; x_{lp})}^{x_l^{'}}.$$

Thus the *basic estimable functions* are $x_l^{'}\boldsymbol{\beta}$. There are $r < n_T$ distinct *basic estimable functions* in one-factor studies, since the rows of the $\boldsymbol{X}$ matrix are identical for subjects belonging to the same treatment. For these distinct *basic estimable functions* $x_i^{'}\boldsymbol{\beta}$, we have that $E\left[Y_i\right] = x_i^{'}\boldsymbol{\beta}$, for $i = 1, \dots, r$.

We can build any estimable function from the *basic* ones, by linear combination of the $x_i^{'}\boldsymbol{\beta}$. Thus, for a

study with a 3 level factor, and unequal sample sizes n_1, n_2, n_3, let us consider the $\boldsymbol{X}$ matrix as defined before:

$$X = \left(\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{array}\right)$$

The *basic estimable functions* are in this case:

$$\begin{aligned} E\left[Y_1\right] &= \mu. + \tau_1 = (1 \; 1 \; 0 \; 0)\boldsymbol{\beta} = x_1^{'}\boldsymbol{\beta}, \\ E\left[Y_2\right] &= \mu. + \tau_2 = (1 \; 0 \; 1 \; 0)\boldsymbol{\beta} = x_2^{'}\boldsymbol{\beta}, \\ E\left[Y_3\right] &= \mu. + \tau_3 = (1 \; 0 \; 0 \; 1)\boldsymbol{\beta} = x_3^{'}\boldsymbol{\beta}. \end{aligned}$$

and any estimable function can be obtained as a linear combination of the three *basic estimable functions*:

$$\begin{aligned} \boldsymbol{q}^{'}\boldsymbol{\beta} &= t_1 x_1^{'}\boldsymbol{\beta} + t_2 x_2^{'}\boldsymbol{\beta} + t_3 x_3^{'}\boldsymbol{\beta} \\ &= t_1\left(\mu. + \tau_1\right) + t_2\left(\mu. + \tau_2\right) + t_3\left(\mu. + \tau_3\right) \\ &= \left(t_1 + t_2 + t_3\right)\mu. + t_1\tau_1 + t_2\tau_2 + t_3\tau_3 \\ &= L_1\mu. + L_2\tau_1 + L_3\tau_2 + \left(L_1 - L_2 - L_3\right)\tau_3, \end{aligned}$$

with $L_1 = t_1 + t_2 + t_3$, $L_2 = t_1$ and $L_3 = t_2$. So, for instance, we can obtain the following estimable functions from this model parametrization:

- $$\left.\begin{array}{lcr} L_1 & = & 0 \\ L_2 & = & 1 \\ L_3 & = & -1 \end{array}\right\} \Rightarrow \quad \boldsymbol{q}^{'}\boldsymbol{\beta} = \tau_1 - \tau_2$$

- $$\left.\begin{array}{lcr} L_1 & = & 1 \\ L_2 & = & 0 \\ L_3 & = & 0 \end{array}\right\} \Rightarrow \quad \boldsymbol{q}^{'}\boldsymbol{\beta} = \mu. + \tau_3$$

- $$\left.\begin{array}{lcr} L_1 & = & 0 \\ L_2 & = & 1 \\ L_3 & = & 1 \end{array}\right\} \Rightarrow \quad \boldsymbol{q}^{'}\boldsymbol{\beta} = \tau_1 + \tau_2 - 2\tau_3$$

But not all linear combinations of the original parameters will be estimable. For instance:

- is $(\tau_1 + \tau_2)$ estimable?

  $$\left.\begin{array}{lcr} L_1 & = & 2 \\ L_2 & = & 1 \\ L_3 & = & 1 \end{array}\right\} \Rightarrow \quad \boldsymbol{q}'\boldsymbol{\beta} = 2\mu. + \tau_1 + \tau_2$$

  It is not possible to obtain coefficients L_1, L_2, L_3 that would give rise to $(\tau_1 + \tau_2)$. Hence, $(\tau_1 + \tau_2)$ is *not estimable.*

This means that the decision on how to set the parameters of the model will influence the further inference that will be possible to perform.

Summarizing

We looked at what happens when we estimate in a particular way an ANOVA model using the regression approach. We can summarize the main ideas as follows:

- we considered the case of an alternative (over-)parametrization of a one-factor model (as an example);
- we saw that, when the model is over-parametrized, we need to "re-parametrize" it in a suitable way (in terms of basic estimable functions);
- we saw that at that point only the quantities that are possible to build as a specific linear combination of the basic estimable functions, are estimable.

The development of theory of estimable functions that we just saw, is at the basis of analysis of ANOVA models in SAS. Transferring these ideas to the use of R for the analysis of ANOVA models, we will drop the concept of estimable functions and work directly with models comparisons. In general, the idea remains the same. Any estimation/testing ANOVA problem within a regression framework can be solved by (as we have already seen):

- parametrizing the model in a suitable way, according to the specific quantities we are interested in,
- knowing how to build the quantity of interest as a linear combination of the parameters of the model,
- or, equivalently, knowing which model comparisons are needed to perform tests on the specific quantities of interest.

2 Type I to III sum of squares

2.1 Different types of tests

Motivation

Let us consider a two-factor study, and let us focus on the regression approach to the analysis. There are in general several possible ways to define the effect of the factors. In principle we might be interested in:

- the sequential effect of factors added in the model (if nested factors are present),
- the main effects only when the interaction effect is not significant (not included in the model),
- the marginal influence of main effects in a model including already the interaction effects...

According to which of these different hypotheses is of interest, we will have to work with different definitions of sum of squares (SS). The three types sums of squares (SS) are all defined as the difference between the error sum of squares (SSE) in two specific nested models that need to be compared. The difference among the three types SS regards the differences among the couples of full and reduced models that are to be compared. In particular we can define:

- *type I SS (sequential testing)*: we test the effect of elements in the model, in the order they appear. Thus each effect is adjusted only for the preceding effects in the model. This kind of testing can be an appropriate choice in nested factors designs;
- *type II SS (hierarchical testing)*: we test the effect of elements in the model, in an order that respects the hierarchy of the effects (hence interaction are tested including main effects in the model, but main effects are tested without considering interaction). Thus at first the interaction effect is adjusted for the main effects in the model, and then possibly the main effects are studied assuming a no-interaction model. This kind of testing can be an appropriate choice for model building;
- *type III SS (marginal testing)*: we test the marginal effect of elements in the model, regardless the order in which they are put in the model, and the hierarchy existing among the effects.

The way of defining these different hypotheses leads to the definition of different nested models to compare, and hence to different SS. Let us consider a two-way ANOVA design, with the factor effects notation: $y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$. The nested full and reduced models that are compared, will differ in terms of the definition of the cell means

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

by dropping a (larger) number of the parameters in the reduced model (than in the full model). These are the full and reduced models considered by the type I to III SS:

	SS		
Effects	*type I*	*type II*	*type III*
Main A	$M_r : \mu_{ij} = \mu_{..}$ $M_f : \mu_{ij} = \mu_{..} + \alpha_i$	$M_r : \mu_{ij} = \mu_{..} + \beta_j$ $M_f : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j$	$M_r : \mu_{ij} = \mu_{..} + \beta_j + (\alpha\beta)_{ij}$ $M_f : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$
Main B	$M_r : \mu_{ij} = \mu_{..} + \alpha_i$ $M_f : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j$	$M_r : \mu_{ij} = \mu_{..} + \alpha_i$ $M_f : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j$	$M_r : \mu_{ij} = \mu_{..} + \alpha_i + (\alpha\beta)_{ij}$ $M_f : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$
Interaction AB	$M_r : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j$ $M_f : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	$M_r : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j$ $M_f : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	$M_r : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j$ $M_f : \mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$

The tests for main and interaction effects, using the several SS are then all built as follows:

$$F^{SS\cdots} = \frac{SSE_r - SSE_f}{\underbrace{df_r - df_f}_{(n_T-p_r)-(n_T-p_f)=p_f-p_r}} \Big/ \frac{SSE_{all}}{\underbrace{df_{all}}_{n_T-ab}},$$

where SSE_r, df_r, p_r and SSE_f, df_f, p_f are respectively the *error sums of squares*, *degrees of freedom*, *number of regression parameters* for the reduced and the full model (from previous slide), and SSE_{all} is the SSE from the complete model with both main effects and interaction. Notice that SSE_{all} is always used at the denominator of the test statistic. The observed value of $F^{SS\cdots}$ needs to be compared with the null distribution $F_{((p_f-p_r,n_T-ab))}$. Large values of $F^{SS\cdots}$ are evidence in favour of H_1. Even though the several SS look quite different, in balanced designs they actually coincide. This is due to the fact that main effects and interactions are orthogonal in balanced designs (thus the total SS can be decomposed in the sum of the single parts). This is not the case for unbalanced designs. These are the relationships between type I to III SS for main and interaction effects, in balanced and unbalanced designs:

	Type of design	
Effects	*Balanced*	*Unbalanced*
Main A	$I = II = III$	$-$
Main B	$I = II = III$	$I = II$
Interaction AB	$I = II = III$	$I = II = III$

2.2 Type I to III SS in R

Test for Type I to III SS in R

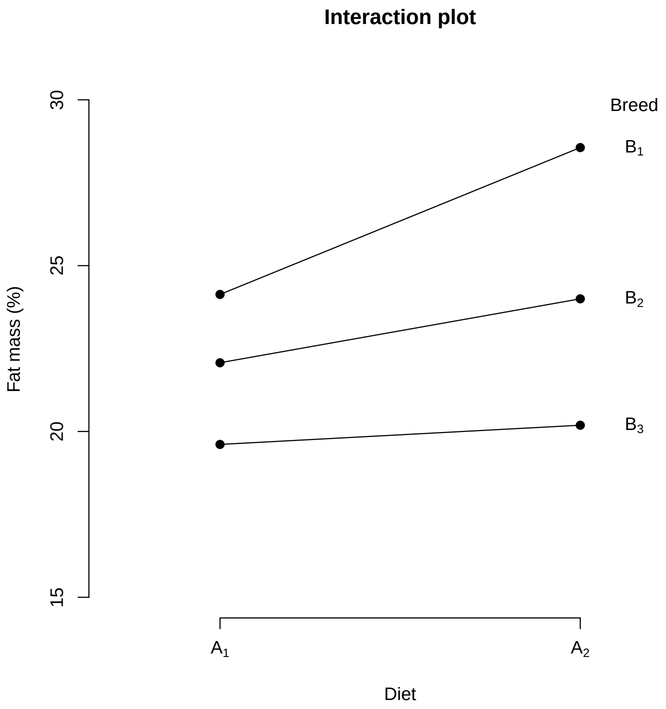
We can easily work with these three types SS in R, in one of the following ways:

- by manually comparing the SSE of the nested full and reduced models (as we saw in lesson 7),
- by working directly on linear combinations of the full model parameters and using the `glht()` function of R (as we saw in lessons 2, 4, 6),

- by using specific built in functions of R, namely `anova()` (as we saw in lessons 1, 4, and 6), and `Anova()`.

Let us see how this works in practice. Let us consider the example on dogs from lessons 5, 6 and 7. Let us start with the unbalanced design considered in lesson 7. Let us compute the I SS:

```
> ### Unbalanced case
> ## Example from lesson 7
> ## Simulating Dogs example
> means_ex3 = array(c(24,28,22,24,20,20),dim=c(2,3))
> n = c(10,7,5,5,7,7)
> set.seed(1)
> fat_11 = rnorm(n[1],means_ex3[1,1],sd=1)
> set.seed(2)
> fat_12 = rnorm(n[2],means_ex3[1,2],sd=1)
> set.seed(3)
> fat_13 = rnorm(n[3],means_ex3[1,3],sd=1)
> set.seed(4)
> fat_21 = rnorm(n[4],means_ex3[2,1],sd=1)
> set.seed(5)
> fat_22 = rnorm(n[5],means_ex3[2,2],sd=1)
> set.seed(6)
> fat_23 = rnorm(n[6],means_ex3[2,3],sd=1)
> fat = c(fat_11,fat_12,fat_13,fat_21,fat_22,fat_23)

> factorA = as.factor(c(rep("D1",22),rep("D2",19)))
> factorB = as.factor(c(rep("Labrador",10),rep("Boxer",7),rep("Hound",5),
+             rep("Labrador",5),rep("Boxer",7),rep("Hound",7)))
> data_dogs = data.frame(cbind("fat"=fat,
+                              "factorA"=factorA,
+                              "factorB"=factorB))
> data_dogs$factorA = factor(factorA,levels=c("D1","D2"))
> data_dogs$factorB = factor(factorB,levels=c("Labrador","Boxer","Hound"))
> contrasts(data_dogs$factorA) = contr.sum
> contrasts(data_dogs$factorB) = contr.sum

> plot(c(1,2),hat_mu[,1],main="Interaction plot",
+      xlab="Diet",ylab="Fat mass (%)",type="l",axes=F,
+      ylim=c(15,max(hat_mu[2,])+2),
+      xlim=c(0.7,2.3))
> points(c(1,2),hat_mu[,1],pch=19)
> points(c(1,2),hat_mu[,2],type="l")
> points(c(1,2),hat_mu[,2],pch=19)
> points(c(1,2),hat_mu[,3],type="l")
> points(c(1,2),hat_mu[,3],pch=19)
> axis(1,at=c(1,2),labels=c(expression(A[1]),
+                           expression(A[2])))
> axis(2)
> text(2+0.15,hat_mu[2,1]+1.3,label="Breed")
> text(2+0.15,hat_mu[2,1],label=expression(B[1]))
> text(2+0.15,hat_mu[2,2],label=expression(B[2]))
> text(2+0.15,hat_mu[2,3],label=expression(B[3]))
```



Let us compute the type I SS:

```
> ## I type SS
> ## by hands: computing SSE_r and SSE_f and performing the test
> # main effect A
> mod_f = lm(fat~factorA*factorB,data=data_dogs)
> summary(mod_f)

Call:
lm(formula = fat ~ factorA * factorB, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4976 -0.7587  0.0602  0.5862  1.7122 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       23.09350    0.14473 159.566  < 2e-16 ***
factorA1          -1.15549    0.14473  -7.984 2.15e-09 ***
factorB1           3.25230    0.20309  16.014  < 2e-16 ***
factorB2          -0.05778    0.20070  -0.288    0.775    
factorA1:factorB1 -1.05811    0.20309  -5.210 8.50e-06 ***
factorA1:factorB2  0.19199    0.20070   0.957    0.345    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9011 on 35 degrees of freedom
Multiple R-squared: 0.9116,	Adjusted R-squared: 0.8989 
F-statistic: 72.15 on 5 and 35 DF,  p-value: < 2.2e-16

> mod_fA_I = lm(fat~factorA,data=data_dogs)
> summary(mod_fA_I)
```

```
Call:
lm(formula = fat ~ factorA, data = data_dogs)

Residuals:
   Min     1Q Median     3Q    Max
-5.104 -2.190 -0.264  1.881  5.840

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.1221     0.4364  52.983   <2e-16 ***
factorA1     -0.6732     0.4364  -1.543    0.131
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.787 on 39 degrees of freedom
Multiple R-squared: 0.0575,     Adjusted R-squared: 0.03334
F-statistic: 2.379 on 1 and 39 DF,  p-value: 0.131

> mod_rA_I = lm(fat~1,data=data_dogs)
> summary(mod_rA_I)

Call:
lm(formula = fat ~ 1, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3820 -2.2042  0.0916  1.4146  6.5628

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.0728     0.4427   52.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.835 on 40 degrees of freedom

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_fA_I = length(coef(mod_fA_I))
> p_rA_I = length(coef(mod_rA_I))

> n_T
[1] 41
> p_f
[1] 6
> p_fA_I
[1] 2
> p_rA_I
[1] 1
> f_regrA_I = ((sum(mod_rA_I$residuals^2)-sum(mod_fA_I$residuals^2))/(p_fA_I-p_rA_I))/
+             (sum(mod_f$residuals^2)/(n_T-p_f))   # we use SSE of the full model
>                                                  # at the denominator of teh F test
> pvalue_factorA_I = 1-pf(f_regrA_I,p_fA_I-p_rA_I,n_T-p_f)
>
```

```
> f_regrA_I
[1] 22.7582
> pvalue_factorA_I
[1] 3.202164e-05

> # main effect B
> mod_fB_I = lm(fat~factorA+factorB,data=data_dogs)
> summary(mod_fB_I)

Call:
lm(formula = fat ~ factorA + factorB, data = data_dogs)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.24743 -0.82785 -0.07679  0.86146  2.44205 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  22.9297     0.1873 122.440  < 2e-16 ***
factorA1     -1.1892     0.1909  -6.230 3.06e-07 ***
factorB1      3.0747     0.2633  11.676 5.71e-14 ***
factorB2      0.1060     0.2625   0.404    0.689    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.192 on 37 degrees of freedom
Multiple R-squared: 0.8364,     Adjusted R-squared: 0.8232 
F-statistic: 63.06 on 3 and 37 DF,  p-value: 1.295e-14

> mod_rB_I = lm(fat~factorA,data=data_dogs)
> summary(mod_rB_I)

Call:
lm(formula = fat ~ factorA, data = data_dogs)

Residuals:
   Min     1Q Median     3Q    Max 
-5.104 -2.190 -0.264  1.881  5.840 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  23.1221     0.4364  52.983   <2e-16 ***
factorA1     -0.6732     0.4364  -1.543    0.131    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.787 on 39 degrees of freedom
Multiple R-squared: 0.0575,     Adjusted R-squared: 0.03334 
F-statistic: 2.379 on 1 and 39 DF,  p-value: 0.131

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_fB_I = length(coef(mod_fB_I))
> p_rB_I = length(coef(mod_rB_I))

> n_T
```

```
[1] 41
> p_f
[1] 6
> p_fB_I
[1] 4
> p_rB_I
[1] 2
> f_regrB_I = ((sum(mod_rB_I$residuals^2)-sum(mod_fB_I$residuals^2))/(p_fB_I-p_rB_I))/
+             (sum(mod_f$residuals^2)/(n_T-p_f))   # we use SSE of the full model
>                                                  # at the denominator of teh F test
> pvalue_factorB_I = 1-pf(f_regrB_I,p_fB_I-p_rB_I,n_T-p_f)
> f_regrB_I
[1] 154.133
> pvalue_factorB_I
[1] 0

> # interaction AB
> mod_fAB_I = lm(fat~factorA+factorB+factorA:factorB,data=data_dogs)
> summary(mod_fAB_I)   # coimcided with mod_f

Call:
lm(formula = fat ~ factorA + factorB + factorA:factorB, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4976 -0.7587  0.0602  0.5862  1.7122

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.09350    0.14473 159.566  < 2e-16 ***
factorA1          -1.15549    0.14473  -7.984 2.15e-09 ***
factorB1           3.25230    0.20309  16.014  < 2e-16 ***
factorB2          -0.05778    0.20070  -0.288    0.775
factorA1:factorB1 -1.05811    0.20309  -5.210 8.50e-06 ***
factorA1:factorB2  0.19199    0.20070   0.957    0.345
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9011 on 35 degrees of freedom
Multiple R-squared: 0.9116,     Adjusted R-squared: 0.8989
F-statistic: 72.15 on 5 and 35 DF,  p-value: < 2.2e-16

> mod_rAB_I = lm(fat~factorA+factorB,data=data_dogs)
> summary(mod_rAB_I)

Call:
lm(formula = fat ~ factorA + factorB, data = data_dogs)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24743 -0.82785 -0.07679  0.86146  2.44205

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.9297     0.1873 122.440  < 2e-16 ***
factorA1     -1.1892     0.1909  -6.230 3.06e-07 ***
```

```
factorB1      3.0747     0.2633  11.676 5.71e-14 ***
factorB2      0.1060     0.2625   0.404    0.689
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.192 on 37 degrees of freedom
Multiple R-squared: 0.8364,     Adjusted R-squared: 0.8232
F-statistic: 63.06 on 3 and 37 DF,  p-value: 1.295e-14

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_fAB_I = length(coef(mod_fAB_I))
> p_rAB_I = length(coef(mod_rAB_I))

> n_T
[1] 41
> p_f
[1] 6
> p_fAB_I
[1] 6
> p_rAB_I
[1] 4
> f_regrAB_I = ((sum(mod_rAB_I$residuals^2)-sum(mod_fAB_I$residuals^2))/(p_fAB_I-p_rAB_I))/
+             (sum(mod_f$residuals^2)/(n_T-p_f))   # we use SSE of the full model
>                                                  # at the denominator of the F test
> pvalue_factorAB_I = 1-pf(f_regrAB_I,p_fAB_I-p_rAB_I,n_T-p_f)
> f_regrAB_I
[1] 14.87044
> pvalue_factorAB_I
[1] 2.116245e-05
> anova(mod_f)   # checking the results (SS and p-values)
Analysis of Variance Table

Response: fat
                Df  Sum Sq Mean Sq F value    Pr(>F)
factorA          1  18.481  18.481  22.758 3.202e-05 ***
factorB          2 250.330 125.165 154.133 < 2.2e-16 ***
factorA:factorB  2  24.151  12.076  14.870 2.116e-05 ***
Residuals       35  28.422   0.812
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Let us compute the type II SS:

```
> ## II type SS
> ## by hands: computing SSE_r and SSE_f and performing the test
> # main effect A
> mod_fA_II = lm(fat~factorA+factorB,data=data_dogs)
> summary(mod_fA_II)

Call:
lm(formula = fat ~ factorA + factorB, data = data_dogs)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24743 -0.82785 -0.07679  0.86146  2.44205
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.9297     0.1873 122.440  < 2e-16 ***
factorA1     -1.1892     0.1909  -6.230 3.06e-07 ***
factorB1      3.0747     0.2633  11.676 5.71e-14 ***
factorB2      0.1060     0.2625   0.404    0.689
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.192 on 37 degrees of freedom
Multiple R-squared: 0.8364,     Adjusted R-squared: 0.8232
F-statistic: 63.06 on 3 and 37 DF,  p-value: 1.295e-14

> mod_rA_II = lm(fat~factorB,data=data_dogs)
> summary(mod_rA_II)

Call:
lm(formula = fat ~ factorB, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4436 -1.1160 -0.2397  0.5521  4.0277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.8636     0.2641  86.571  < 2e-16 ***
factorB1      2.7443     0.3643   7.532 4.74e-09 ***
factorB2      0.1721     0.3705   0.465    0.645
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.684 on 38 degrees of freedom
Multiple R-squared: 0.6648,     Adjusted R-squared: 0.6472
F-statistic: 37.68 on 2 and 38 DF,  p-value: 9.569e-10

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_fA_II = length(coef(mod_fA_II))
> p_rA_II = length(coef(mod_rA_II))

> n_T
[1] 41
> p_f
[1] 6
> p_fA_II
[1] 4
> p_rA_II
[1] 3
> f_regrA_II = ((sum(mod_rA_II$residuals^2)-sum(mod_fA_II$residuals^2))/(p_fA_II-p_rA_II))/
+             (sum(mod_f$residuals^2)/(n_T-p_f))   # we use SSE of the full model
>                                                  # at the denominator of the F test
> pvalue_factorA_II = 1-pf(f_regrA_II,p_fA_II-p_rA_II,n_T-p_f)
>
> f_regrA_II
```

```
[1] 67.92114
> pvalue_factorA_II
[1] 1.031698e-09

> # main effect B
> mod_fB_II = lm(fat~factorA+factorB,data=data_dogs)
> summary(mod_fA_II)

Call:
lm(formula = fat ~ factorA + factorB, data = data_dogs)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.24743 -0.82785 -0.07679  0.86146  2.44205 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  22.9297     0.1873 122.440  < 2e-16 ***
factorA1     -1.1892     0.1909  -6.230 3.06e-07 ***
factorB1      3.0747     0.2633  11.676 5.71e-14 ***
factorB2      0.1060     0.2625   0.404    0.689    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.192 on 37 degrees of freedom
Multiple R-squared: 0.8364,     Adjusted R-squared: 0.8232 
F-statistic: 63.06 on 3 and 37 DF,  p-value: 1.295e-14

> mod_rB_II = lm(fat~factorA,data=data_dogs)
> summary(mod_rB_II)

Call:
lm(formula = fat ~ factorA, data = data_dogs)

Residuals:
   Min     1Q Median     3Q    Max 
-5.104 -2.190 -0.264  1.881  5.840 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  23.1221     0.4364  52.983   <2e-16 ***
factorA1     -0.6732     0.4364  -1.543    0.131    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.787 on 39 degrees of freedom
Multiple R-squared: 0.0575,     Adjusted R-squared: 0.03334 
F-statistic: 2.379 on 1 and 39 DF,  p-value: 0.131

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_fB_II = length(coef(mod_fB_II))
> p_rB_II = length(coef(mod_rB_II))

> n_T
[1] 41
```

```
> p_f
[1] 6
> p_fB_II
[1] 4
> p_rB_II
[1] 2
> f_regrB_II = ((sum(mod_rB_II$residuals^2)-sum(mod_fB_II$residuals^2))/(p_fB_II-p_rB_II))/
+             (sum(mod_f$residuals^2)/(n_T-p_f))   # we use SSE of the full model
>                                                  # at the denominator of teh F test
> pvalue_factorB_II = 1-pf(f_regrB_II,p_fB_II-p_rB_II,n_T-p_f)
> f_regrB_II
[1] 154.133
> pvalue_factorB_II
[1] 0

> # interaction AB
> mod_fAB_II = lm(fat~factorA+factorB+factorA:factorB,data=data_dogs)
> summary(mod_fAB_II)   # coimcided with mod_f

Call:
lm(formula = fat ~ factorA + factorB + factorA:factorB, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4976 -0.7587  0.0602  0.5862  1.7122

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.09350    0.14473 159.566  < 2e-16 ***
factorA1          -1.15549    0.14473  -7.984 2.15e-09 ***
factorB1           3.25230    0.20309  16.014  < 2e-16 ***
factorB2          -0.05778    0.20070  -0.288    0.775
factorA1:factorB1 -1.05811    0.20309  -5.210 8.50e-06 ***
factorA1:factorB2  0.19199    0.20070   0.957    0.345
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9011 on 35 degrees of freedom
Multiple R-squared: 0.9116,     Adjusted R-squared: 0.8989
F-statistic: 72.15 on 5 and 35 DF,  p-value: < 2.2e-16

> mod_rAB_II = lm(fat~factorA+factorB,data=data_dogs)
> summary(mod_rAB_II)

Call:
lm(formula = fat ~ factorA + factorB, data = data_dogs)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24743 -0.82785 -0.07679  0.86146  2.44205

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.9297     0.1873 122.440  < 2e-16 ***
factorA1     -1.1892     0.1909  -6.230 3.06e-07 ***
factorB1      3.0747     0.2633  11.676 5.71e-14 ***
```

```
factorB2      0.1060     0.2625   0.404    0.689
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.192 on 37 degrees of freedom
Multiple R-squared: 0.8364,     Adjusted R-squared: 0.8232
F-statistic: 63.06 on 3 and 37 DF,  p-value: 1.295e-14

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_fAB_II = length(coef(mod_fAB_II))
> p_rAB_II = length(coef(mod_rAB_II))

> n_T
[1] 41
> p_f
[1] 6
> p_fAB_II
[1] 6
> p_rAB_II
[1] 4
> f_regrAB_II = ((sum(mod_rAB_II$residuals^2)-sum(mod_fAB_II$residuals^2))/(p_fAB_II-p_rAB_II))/
+             (sum(mod_f$residuals^2)/(n_T-p_f))   # we use SSE of the full model
>                                                  # at the denominator of the F test
> pvalue_factorAB_II = 1-pf(f_regrAB_II,p_fAB_II-p_rAB_II,n_T-p_f)
> f_regrAB_II
[1] 14.87044
> pvalue_factorAB_II
[1] 2.116245e-05
> Anova(mod_f,type="II")
Anova Table (Type II tests)

Response: fat
                 Sum Sq Df F value    Pr(>F)
factorA          55.156  1  67.921 1.032e-09 ***
factorB         250.330  2 154.133 < 2.2e-16 ***
factorA:factorB  24.151  2  14.870 2.116e-05 ***
Residuals        28.422 35
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Let us compute the type III SS:

```
> ## III type SS
> ## by hands: computing SSE_r and SSE_f and performing the test
> # main effect A
> mod_fA_III = lm(fat~factorA+factorB+factorA:factorB,data=data_dogs)
> summary(mod_fA_III)

Call:
lm(formula = fat ~ factorA + factorB + factorA:factorB, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4976 -0.7587  0.0602  0.5862  1.7122
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       23.09350    0.14473 159.566  < 2e-16 ***
factorA1          -1.15549    0.14473  -7.984 2.15e-09 ***
factorB1           3.25230    0.20309  16.014  < 2e-16 ***
factorB2          -0.05778    0.20070  -0.288    0.775    
factorA1:factorB1 -1.05811    0.20309  -5.210 8.50e-06 ***
factorA1:factorB2  0.19199    0.20070   0.957    0.345    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9011 on 35 degrees of freedom
Multiple R-squared: 0.9116,     Adjusted R-squared: 0.8989 
F-statistic: 72.15 on 5 and 35 DF,  p-value: < 2.2e-16

> X0 <- model.matrix(mod_f)
> X1 <- X0[,!colnames(X0) %in% "factorA1"]  # we need to modify the
>                                           # matrix of regressors by hand
>                                           # (R tends to reparametrize the
>                                           #  model otherwise)
> mod_rA_III = lm(fat~0+X1,data=data_dogs)
> summary(mod_rA_III)

Call:
lm(formula = fat ~ 0 + X1, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2692 -1.0064 -0.1881  1.1375  2.7788 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
X1(Intercept)       23.04017    0.23944  96.227  < 2e-16 ***
X1factorB1           2.93232    0.32974   8.893  1.3e-10 ***
X1factorB2          -0.00445    0.33221  -0.013  0.98939    
X1factorA1:factorB1 -1.09366    0.33628  -3.252  0.00249 ** 
X1factorA1:factorB2  0.10311    0.33188   0.311  0.75783    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.492 on 36 degrees of freedom
Multiple R-squared: 0.9964,     Adjusted R-squared: 0.9959 
F-statistic:  1982 on 5 and 36 DF,  p-value: < 2.2e-16

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_fA_III = length(coef(mod_fA_III))
> p_rA_III = length(coef(mod_rA_III))
> n_T
[1] 41
> p_f
[1] 6
> p_fA_III
[1] 6
> p_rA_III
```

```
[1] 5
> f_regrA_III = ((sum(mod_rA_III$residuals^2)-sum(mod_fA_III$residuals^2))/(p_fA_III-p_rA_III))/
+              (sum(mod_f$residuals^2)/(n_T-p_f))   # we use SSE of the full model
>                                                   # at the denominator of teh F test
> pvalue_factorA_III = 1-pf(f_regrA_III,p_fA_III-p_rA_III,n_T-p_f)
>
> f_regrA_III
[1] 63.7433
> pvalue_factorA_III
[1] 2.152357e-09

> # main effect B
> mod_fB_III = lm(fat~factorA+factorB+factorA:factorB,data=data_dogs)
> summary(mod_fB_III)

Call:
lm(formula = fat ~ factorA + factorB + factorA:factorB, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4976 -0.7587  0.0602  0.5862  1.7122

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.09350    0.14473 159.566  < 2e-16 ***
factorA1          -1.15549    0.14473  -7.984 2.15e-09 ***
factorB1           3.25230    0.20309  16.014  < 2e-16 ***
factorB2          -0.05778    0.20070  -0.288    0.775
factorA1:factorB1 -1.05811    0.20309  -5.210 8.50e-06 ***
factorA1:factorB2  0.19199    0.20070   0.957    0.345
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9011 on 35 degrees of freedom
Multiple R-squared: 0.9116,     Adjusted R-squared: 0.8989
F-statistic: 72.15 on 5 and 35 DF,  p-value: < 2.2e-16

> X0 <- model.matrix(mod_f)
> X1 <- X0[,(!colnames(X0) %in% "factorB1") & (!colnames(X0) %in% "factorB2")]  # we need to modify
>                                            # matrix of regressors by hand
>                                            # (R tends to reparametrize the
>                                            #  model otherwise)
> mod_rB_III = lm(fat~0+X1,data=data_dogs)
> summary(mod_rB_III)

Call:
lm(formula = fat ~ 0 + X1, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6468 -2.0947 -0.1224  2.1293  5.2319

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
X1(Intercept)        23.2290     0.4466  52.010   <2e-16 ***
X1factorA1           -0.6242     0.4382  -1.424    0.163
```

```
X1factorA1:factorB1  -0.5505     0.6162  -0.893    0.377
X1factorA1:factorB2  -0.3393     0.6142  -0.552    0.584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.789 on 37 degrees of freedom
Multiple R-squared: 0.987,      Adjusted R-squared: 0.9856
F-statistic: 702.5 on 4 and 37 DF,  p-value: < 2.2e-16

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_fB_III = length(coef(mod_fB_III))
> p_rB_III = length(coef(mod_rB_III))
> n_T
[1] 41
> p_f
[1] 6
> p_fB_III
[1] 6
> p_rB_III
[1] 4
> f_regrB_III = ((sum(mod_rB_III$residuals^2)-sum(mod_fB_III$residuals^2))/(p_fB_III-p_rB_III))/
+             (sum(mod_f$residuals^2)/(n_T-p_f))   # we use SSE of the full model
>                                                  # at the denominator of teh F test
> pvalue_factorB_III = 1-pf(f_regrB_III,p_fB_III-p_rB_III,n_T-p_f)
> f_regrB_III
[1] 159.7309
> pvalue_factorB_III
[1] 0

> # interaction AB
> mod_fAB_III = lm(fat~factorA+factorB+factorA:factorB,data=data_dogs)
> summary(mod_fAB_III)   # coimcided with mod_f

Call:
lm(formula = fat ~ factorA + factorB + factorA:factorB, data = data_dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4976 -0.7587  0.0602  0.5862  1.7122

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.09350    0.14473 159.566  < 2e-16 ***
factorA1          -1.15549    0.14473  -7.984 2.15e-09 ***
factorB1           3.25230    0.20309  16.014  < 2e-16 ***
factorB2          -0.05778    0.20070  -0.288    0.775
factorA1:factorB1 -1.05811    0.20309  -5.210 8.50e-06 ***
factorA1:factorB2  0.19199    0.20070   0.957    0.345
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9011 on 35 degrees of freedom
Multiple R-squared: 0.9116,     Adjusted R-squared: 0.8989
F-statistic: 72.15 on 5 and 35 DF,  p-value: < 2.2e-16
```

```
> mod_rAB_III = lm(fat~factorA+factorB,data=data_dogs)
> summary(mod_rAB_III)

Call:
lm(formula = fat ~ factorA + factorB, data = data_dogs)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.24743 -0.82785 -0.07679  0.86146  2.44205 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  22.9297     0.1873 122.440  < 2e-16 ***
factorA1     -1.1892     0.1909  -6.230 3.06e-07 ***
factorB1      3.0747     0.2633  11.676 5.71e-14 ***
factorB2      0.1060     0.2625   0.404    0.689    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.192 on 37 degrees of freedom
Multiple R-squared: 0.8364,     Adjusted R-squared: 0.8232 
F-statistic: 63.06 on 3 and 37 DF,  p-value: 1.295e-14

> # perform the anova test
> n_T = dim(data_dogs)[1]
> p_f = length(coef(mod_f))
> p_fAB_III = length(coef(mod_fAB_III))
> p_rAB_III = length(coef(mod_rAB_III))

> n_T
[1] 41
> p_f
[1] 6
> p_fAB_III
[1] 6
> p_rAB_III
[1] 4
> f_regrAB_III = ((sum(mod_rAB_III$residuals^2)-sum(mod_fAB_III$residuals^2))/
                 (p_fAB_III-p_rAB_III))/
+              (sum(mod_f$residuals^2)/(n_T-p_f))   # we use SSE of the full model
>                                                   # at the denominator of the F test
> pvalue_factorAB_III = 1-pf(f_regrAB_II,p_fAB_II-p_rAB_II,n_T-p_f)
> f_regrAB_III
[1] 14.87044
> pvalue_factorAB_III
[1] 2.116245e-05
> Anova(mod_f,type="III")
Anova Table (Type III tests)

Response: fat
                 Sum Sq Df   F value    Pr(>F)    
(Intercept)     20676.0  1 25461.240 < 2.2e-16 ***
factorA            51.8  1    63.743 2.152e-09 ***
factorB           259.4  2   159.731 < 2.2e-16 ***
factorA:factorB    24.2  2    14.870 2.116e-05 ***
Residuals          28.4 35                        
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The type III SS for main effect of factor A, can be computed also using the function `glht()` in R. We need to select the related coefficient when defining the contrasts to be tested:

```
> # III type SS, example using the function glht()
> gA=glht(mod_f,linfct=rbind(c(0,1,0,0,0,0)))
> summary(gA)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = fat ~ factorA * factorB, data = data_dogs)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0  -1.1555     0.1447  -7.984 2.15e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

At last, let us check that the three types of sums of squares do not change for balanced designs. Let us consider the example on dogs from lesson 5 and 6:

```
> ### Balanced case
> ## Building the dataset
> ## (simulating data with interaction,
> ## like in example 3 of lesson 5)
> means_ex3 = array(c(24,28,22,24,20,20),dim=c(2,3))
> n = 10
> set.seed(1)
> fat_11 = rnorm(n,means_ex3[1,1],sd=1)
> set.seed(2)
> fat_12 = rnorm(n,means_ex3[1,2],sd=1)
> set.seed(3)
> fat_13 = rnorm(n,means_ex3[1,3],sd=1)
> set.seed(4)
> fat_21 = rnorm(n,means_ex3[2,1],sd=1)
> set.seed(5)
> fat_22 = rnorm(n,means_ex3[2,2],sd=1)
> set.seed(6)
> fat_23 = rnorm(n,means_ex3[2,3],sd=1)
> fat = c(fat_11,fat_12,fat_13,fat_21,fat_22,fat_23)

> factorA = as.factor(c(rep("D1",3*n),rep("D2",3*n)))
> factorB = as.factor(c(rep("Labrador",n),rep("Boxer",n),rep("Hound",n),
+             rep("Labrador",n),rep("Boxer",n),rep("Hound",n)))
> data_dogs_balanced = data.frame(cbind("fat"=fat,
+                               "factorA"=factorA,
+                               "factorB"=factorB))
> data_dogs_balanced$factorA = factor(factorA,levels=c("D1","D2"))
> data_dogs_balanced$factorB = factor(factorB,levels=c("Labrador","Boxer","Hound"))
> str(data_dogs_balanced)
'data.frame':   60 obs. of  3 variables:
 $ fat    : num  23.4 24.2 23.2 25.6 24.3 ...
 $ factorA: Factor w/ 2 levels "D1","D2": 1 1 1 1 1 1 1 1 1 1 ...
 $ factorB: Factor w/ 3 levels "Labrador","Boxer",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
> ## Estimating the models and performing the F tests:
> # define the contrasts
> contrasts(data_dogs_balanced$factorA) = contr.sum
> contrasts(data_dogs_balanced$factorB) = contr.sum
> # fit the initial model
> mod1_dogs_ex3 = lm(fat~factorA*factorB,data=data_dogs_balanced)
> summary(mod1_dogs_ex3)

Call:
lm(formula = fat ~ factorA * factorB, data = data_dogs_balanced)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84778 -0.74117  0.00157  0.45688  1.79029

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.14488    0.11991 193.012  < 2e-16 ***
factorA1          -1.05280    0.11991  -8.780 5.55e-12 ***
factorB1           3.20449    0.16958  18.896  < 2e-16 ***
factorB2          -0.07873    0.16958  -0.464    0.644
factorA1:factorB1 -1.16436    0.16958  -6.866 6.80e-09 ***
factorA1:factorB2  0.19780    0.16958   1.166    0.249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9288 on 54 degrees of freedom
Multiple R-squared: 0.9169,     Adjusted R-squared: 0.9092
F-statistic: 119.2 on 5 and 54 DF,  p-value: < 2.2e-16

> # perform the anova tests
> anova(mod1_dogs_ex3)
Analysis of Variance Table

Response: fat
                Df Sum Sq Mean Sq F value    Pr(>F)
factorA          1  66.50  66.504  77.082 5.555e-12 ***
factorB          2 400.91 200.454 232.340 < 2.2e-16 ***
factorA:factorB  2  46.58  23.291  26.996 7.466e-09 ***
Residuals       54  46.59   0.863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> Anova(mod1_dogs_ex3,type="II")
Anova Table (Type II tests)

Response: fat
                Sum Sq Df F value    Pr(>F)
factorA          66.50  1  77.082 5.555e-12 ***
factorB         400.91  2 232.340 < 2.2e-16 ***
factorA:factorB  46.58  2  26.996 7.466e-09 ***
Residuals        46.59 54
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Anova(mod1_dogs_ex3,type="III")   # for balanced designs
Anova Table (Type III tests)
```

```
Response: fat
                Sum Sq Df   F value    Pr(>F)    
(Intercept)      32141  1 37253.825 < 2.2e-16 ***
factorA             67  1    77.082 5.555e-12 ***
factorB            401  2   232.340 < 2.2e-16 ***
factorA:factorB     47  2    26.996 7.466e-09 ***
Residuals           47 54                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```