2.5 ANOVA

Analysis of Variance – ANOVA

Analysis of variance is a method to *compare means of groups of continuous observations* where the groups are defined by the levels of the factors.

- $\mathbf{Y}$: continuous variable
- $\mathbf{x}$: categorical variable(s)

The element of $\mathbf{X}$ (design matrix) are **dummy** variables.

Example: One-factor analysis

Genetically similar seeds are randomly assigned to be raised in either nutritionally enriched environment (treatment A or treatment B) or standard conditions (control group) using a **completely randomised experimental design**. After a predetermined time all plants are harvested, dried and weighted.

```
library(dobson)
data("plant.dried")
attach(plant.dried)

head(plant.dried)
```

Remark: if experimental units are randomly allocated to groups corresponding to J levels of a factor, this is called a **completely randomised experiment**.

The responses at level j, i.e. $Y_{j1}, \ldots, Y_{jn_j}$ are called **replicates**. In the example $n_j = K$ for $j = 1, 2, 3$, but this is not always true. We assume $n_j = K$ throughout for simplicity.

The **response** vector (of length $N = JK$), is given by:

$$\mathbf{y} = [\mathrm{Y}_{11}, \mathrm{Y}_{12}, \ldots, \mathrm{Y}_{1K}, \mathrm{Y}_{21}, \ldots, \mathrm{Y}_{2K}, \ldots, \mathrm{Y}_{J1}, \ldots, \mathrm{Y}_{JK}]$$

We consider **three specifications** of the model:

1. $\mathbb{E}(\mathrm{Y}_{jk}) = \mu_j$ for $k = 1, \ldots, K$
2. $\mathbb{E}(\mathrm{Y}_{jk}) = \mu + \alpha_j$ for $k = 1, \ldots, K$
3. $\mathbb{E}(\mathrm{Y}_{jk}) = \mu + \alpha_j$ for $k = 1, \ldots, K$, under constraint $\alpha_1 = 0$

Model 1

Model 1 - $\mathbb{E}(\mathrm{Y}_{jk}) = \mu_j$ for $k = 1, \ldots, K$

Model (1) can be re-written as $\mathbb{E}(Y_i) = \sum_{j=1}^{J} x_{ij}\mu_j$ for $i = 1, \ldots, N$ where x_{ij} represent an element of the **design matrix** through:

- $x_{ij} = 1$ if response Y_i corresponds to level j
- $x_{ij} = 0$ otherwise

This gives $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ where

$$\boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{pmatrix}.$$

The design matrix is given by

```
X <- cbind( c(rep(1,10),rep(0,20)) , c(rep(0,10),rep(1,10), rep(0,10)) ,
            c(rep(0,20),rep(1,10))  )
```

and the estimate $\hat{\boldsymbol{\beta}}$ is the vector of sample means for each group

```
library(dobson)
data("plant.dried")
attach(plant.dried)

X <- cbind( c(rep(1,10),rep(0,20)) , c(rep(0,10),rep(1,10), rep(0,10)) ,
            c(rep(0,20),rep(1,10))  )

y <- matrix(weight,ncol=1)
b.hat <- solve(t(X) %*% X) %*% t(X) %*% y
b.hat
```

The disadvantage of this simple formulation of the model is that it *cannot be extended to consider more than one factor*.

Model 2

Model 2 - $\mathbb{E}(\mathrm{Y}_{jk}) = \mu + \alpha_j$ for $k = 1, \ldots, K$

In this model μ is an **average effect** for all levels and α_j is an **additional effect** due to level j. In this case we have:

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_J \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{pmatrix}.$$

The design matrix as an additional column of elements equal to 1.

The first row (or column) of the $(J+1) \times (J+1)$ matrix $\mathbf{X}^\top \mathbf{X}$ is the sum of the remaining rows (or columns), therefore $\mathbf{X}^\top \mathbf{X}$ is **singular** and there is **no unique solution**.

The general solution can be written

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_J \end{bmatrix} = \frac{1}{K} \begin{bmatrix} 0 \\ \mathrm{Y}_{1.} \\ \vdots \\ \mathrm{Y}_{J.} \end{bmatrix} - \lambda \begin{bmatrix} -1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

where λ is an arbitrary constant and $\mathrm{Y}_{j.} = \sum_{i=1}^{n_j} \mathrm{Y}_{ij}$.

Usually a **sum-to-one constraint** is used, such that $\sum_{j=1}^{J} \alpha_j = 0$, i.e.

$$\frac{1}{K} \sum_{j=1}^{J} Y_{j.} - J\lambda = 0 \qquad \Longleftrightarrow \qquad \lambda = \frac{1}{JK} \sum_{j=1}^{J} Y_{j.} = \frac{Y_{..}}{N}$$

and therefore

$$\hat{\mu} = \frac{Y_{..}}{N} \qquad \text{and} \qquad \hat{\alpha}_j = \frac{Y_{j.}}{K} - \frac{Y_{..}}{N} \quad j = 1, \ldots, J$$

```
library(dobson)
data("plant.dried")
attach(plant.dried)
```

```
X <- cbind( rep(1, nrow(plant.dried)) ,
            c(rep(1,10),rep(0,20)) ,
            c(rep(0,10),rep(1,10), rep(0,10)) ,
            c(rep(0,20),rep(1,10))  )

lambda <- mean(weight)
mu.hat <- lambda
mu.hat

alpha.hat <- aggregate(weight, list(group), mean)$x - lambda
alpha.hat
```

Model 3

Model 3 - $\mathbb{E}(\mathrm{Y}_{jk}) = \mu + \alpha_j$ for $k = 1, \ldots, K$, under constraint $\alpha_1 = 0$

In this case μ represents the *effect of the first level* and α_j measures *the difference between the first level and the j-th level* of the factor.

This is called a **corner point parametrization**. We have

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_J \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{pmatrix}.$$

Now $\mathbf{X}^\top \mathbf{X}$ is **not singular**, so there a **unique solution** for

$$\hat{\boldsymbol{\beta}} = \frac{1}{K} \begin{bmatrix} \mathrm{Y}_{1.} \\ \mathrm{Y}_{2.} - \mathrm{Y}_{1.} \\ \vdots \\ \mathrm{Y}_{J.} - \mathrm{Y}_{1.} \end{bmatrix}$$

```
library(dobson)
data("plant.dried")
attach(plant.dried)

y <- matrix(weight, ncol=1)
X <- cbind( rep(1, nrow(plant.dried)) ,
            c(rep(0,10),rep(1,10), rep(0,10)) ,
            c(rep(0,20),rep(1,10))  )

b.hat <- solve(t(X) %*% X) %*% t(X) %*% y
b.hat
```

In the analysis of the variance, it is important to *compare the alternative hypothesis* (means for each level differ) with the *null hypothesis* (means are all equal)

For the null model, $\mathbb{E}(\mathrm{Y}_{jk}) = \mu$ and the design matrix is a column vector of elements equal to 1, i.e. $\mathbf{X}^\top \mathbf{X} = N$ and $\mathbf{X}^\top \mathbf{y} = \mathrm{Y}_{..}$.

$$D_1 = \frac{1}{\sigma^2}(\mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{y})$$

$$D_0 = \frac{1}{\sigma^2}\left[\sum_{j=1}^{J}\sum_{k=1}^{K} \mathrm{Y}_{jk}^2 - \frac{Y_{\cdot\cdot}^2}{N}\right]$$

and the F-statistic

$$F = \frac{D_0 - D_1}{J-1} \Big/ \frac{D_1}{N-J}$$

```
library(dobson)
data("plant.dried")
attach(plant.dried)

y <- matrix(weight, ncol=1)
X <- cbind( rep(1, nrow(plants)) ,
            c(rep(0,10),rep(1,10), rep(0,10)) ,
            c(rep(0,20),rep(1,10))  )

b.hat <- solve(t(X) %*% X) %*% t(X) %*% y
b.hat

D1 <- t(y) %*% y - t(b.hat) %*% t(X) %*% y

# Null model:
X0 <- matrix(rep(1,nrow(plant.dried)),ncol=1)
b0.hat <- solve(t(X0) %*% X0) %*% t(X0) %*% y
D0 <- t(y) %*% y - t(b0.hat) %*% t(X0) %*% y

Fstat <- ((D0 - D1)/(ncol(X)-1)) / (D1 / (nrow(X)-ncol(X)))
crit.val <- qf(0.95, df1 = ncol(X)-1, df2 = nrow(X)-ncol(X))

if(Fstat > crit.val){
  cat("There is enough evidence to reject H0")
}else{
  cat("There is NOT enough evidence to reject H0")
}
```

Similarly, we could have obtained the same results with the `lm()` function

```
library(dobson)
data("plant.dried")
attach(plant.dried)

res.lm <- lm(weight ~ group)
summary(res.lm)
```

Have a look to the two-factor analysis of variance, which is a simple extension of this section (Dobson and Barnett, 2018, Section 6.4.2).