Normal Linear Models

Bertrand Servin

`bertrand.servin@inrae.fr`

UE Méthodes Numériques 2023-2024

Outline of the class

1. Introduction

2. Linear Regression
   - Model specification
   - Estimation
   - Goodness of fit
   - Inference on the model parameters

3. ANalysis Of VAriance (ANOVA)
   - One factor ANOVA
   - 2-way ANOVA

Outline

1. Introduction
2. Linear Regression
   - Model specification
   - Estimation
   - Goodness of fit
   - Inference on the model parameters
3. ANalysis Of VAriance (ANOVA)
   - One factor ANOVA
   - 2-way ANOVA

Definitions

Statistical Model A statistical model is a mean to describe the relationships between **response variables (Y)** and **explanatory variables (X)**. A model is a formal representation of a theory (set of hypotheses).

Linear Models A **linear** model is a statistical model where Y are modelled as a linear combination of the model parameters :

$$\mathbf{Y} = \mathbf{X_1}.\beta_\mathbf{1} + \mathbf{X_2}.\beta_\mathbf{2} + ... + \mathbf{e}$$

$$\mathbf{Y} = \mathbf{E}(\mathbf{Y}) + \mathbf{e}$$

Effets The model parameters $\beta_1, \beta_2, \ldots$ are the effects of the explanatory variables.

Résidus Part of the responses **Y** is not explained by the predictors **X** : the errors **e** of the model $(\mathbf{E}(\mathbf{e}) = \mathbf{0})$.

Classes of Linear Models

Response	Predictors	Model
Continuous	Binary	t-test
Unique	Categorical	ANOVA
	Continue	Regression (can be multiple)
	Categorical + continue	multiple regression

General Approach to statistical modeling

Can be decomposed as :

1. **Model specification** : setting up hypotheses, sometimes following initial explanatory data analyses (*e.g.* PCA)
2. **Parameter Estimation** using statistical analysis software (*e.g.* R)
3. Check model fit to the data
4. Statistical inference : hypothesis testing, confidence intervals ...

Outline

1 Introduction

2 Linear Regression

- Model specification
- Estimation
- Goodness of fit
- Inference on the model parameters

3 ANalysis Of VAriance (ANOVA)

- One factor ANOVA
- 2-way ANOVA

Outline

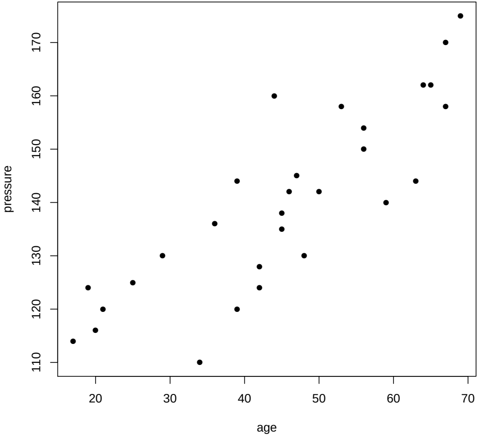
1. Introduction

2. Linear Regression
   - Model specification
   - Estimation
   - Goodness of fit
   - Inference on the model parameters

3. ANalysis Of VAriance (ANOVA)
   - One factor ANOVA
   - 2-way ANOVA

Example : blood pressure and age

- $n = 29$ data points (samples)
- n measures of blood pressure $y_1, \ldots, y_i, \ldots, y_n$
- n values $x_1, \ldots, x_i, \ldots, x_n$ for the age at measurement



Linear relationship between y et x

The general equation of a line is :

$$y \approx \beta_1 + \beta_2 x$$

leads to the simple regression model :

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

we want the errors e_i to be as small as possible (minimize $Var(e_i) = \sigma^2$).

Outline

1. Introduction
2. Linear Regression
   - Model specification
   - Estimation
   - Goodness of fit
   - Inference on the model parameters
3. ANalysis Of VAriance (ANOVA)
   - One factor ANOVA
   - 2-way ANOVA

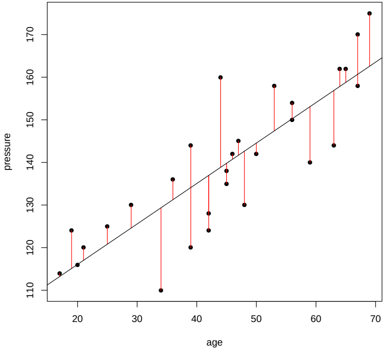
Parameter estimation : least squares

- Compute estimates $\hat{\beta}$ of parameters s β That minimize the sum of squared differences between observations and fitted values :

$$\min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$

- The estimate $\hat{\sigma}^2$ of σ^2 is simply

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2}{n-2}$$

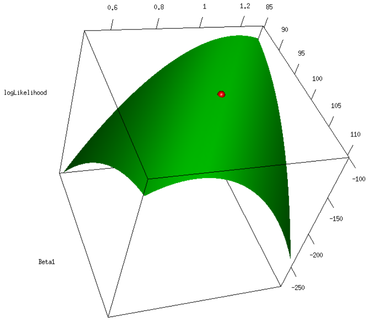


Parameter estimation : maximum likelihood

- Principle : Find parameter values that make the observatons the most likely. Requires to speicify the joint pdf of the observations
- $\hat{\beta}$ and $\hat{\sigma}^2$ such that $\max(\log) f(y; \beta, \sigma^2)$ seen as a function of the parameters :

$$\max L(\beta, \sigma^2; y)$$

  L is the likelihood function.
- It can be shown that the least square estimates of β is the same as the Maximum Likelihood Estimates (MLE)



$$f(y_i; \beta_1, \beta_2, \sigma^2) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2 x_i)^2\right\}$$

Results of model fit

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

Estimates of the parameters $\hat{\beta}$, $\hat{\sigma}^2$ of β and σ^2 resp.

fitted values of the observations $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Residuals (calculated) $\hat{e}_i = y_i - \hat{y}_i$

All of these are only function of the data y_i and x_i

Results of model fit

Estimates of the parameters $\hat{\beta}$, $\hat{\sigma}^2$ of β and σ^2 resp.

fitted values of the observations $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Residuals (calculated) $\hat{e}_i = y_i - \hat{y}_i$

All of these are only function of the data y_i and x_i

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

$$\hat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$$

Results of model fit

Estimates of the parameters $\hat{\beta}$, $\hat{\sigma}^2$ of β and σ^2 resp.

fitted values of the observations $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Residuals (calculated) $\hat{e}_i = y_i - \hat{y}_i$

All of these are only function of the data y_i and x_i

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

$$\hat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$$

$$\widehat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$$

Results of model fit

Estimates of the parameters $\hat{\beta}$, $\hat{\sigma}^2$ of β and σ^2 resp.

fitted values of the observations $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Residuals (calculated) $\hat{e}_i = y_i - \hat{y}_i$

All of these are only function of the data y_i and x_i

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

$$\hat{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

$$\widehat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

$$\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

$\mathbf{H}$ Hat matrix (puts a hat on Y)

Results of model fit

Estimates of the parameters $\hat{\beta}$, $\hat{\sigma}^2$ of β and σ^2 resp.

fitted values of the observations $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Residuals (calculated) $\hat{e}_i = y_i - \hat{y}_i$

All of these are only function of the data y_i and x_i

The matrix $\mathbf{X}^\mathbf{T}\mathbf{X}$ must be invertible.

When the number of covariates is very large (multiple regression), colinearity beween covriates can lead to un-invertible matrices. This leads to problems in parameter estimates (invalid values, very large standard errors) ...

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

$$\hat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$$

$$\widehat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$$

$$\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

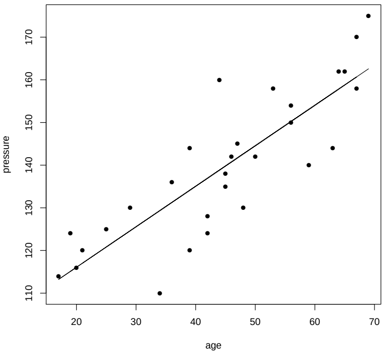
$\mathbf{H}$ Hat matrix (puts a hat on Y)

Fitting a linear regression with R

```
> reg=lm(pressure~age)
> reg

Call:
lm(formula = pressure ~ age)

Coefficients:
(Intercept)          age
    97.0771       0.9493
```



Outline

1. Introduction
2. Linear Regression
   - Model specification
   - Estimation
   - Goodness of fit
   - Inference on the model parameters
3. ANalysis Of VAriance (ANOVA)
   - One factor ANOVA
   - 2-way ANOVA

Criteria

- Inference is only valid if the linear model hypotheses hold :
  - Errors are Gaussian and homogeneous
  - Absence of outliers = datapoints that lie outside of the typical range of values and have a strong influence on the results
- Before making inference on the model parameters, goodness of fit of the model must be assessed

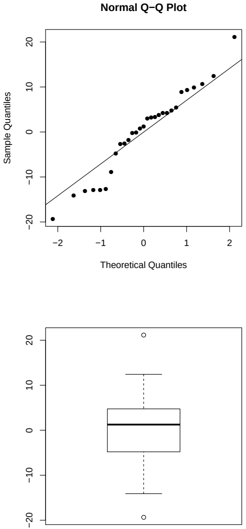
Errors are Gaussian

For a normal linear model, the error (stochastic part) is assumed to be normaly distributed

$$e_i \sim \mathcal{N}(0, \sigma^2)$$

It is also approximately true for the residuals (predictions of the errors $\hat{e}$).

```
par(mfrow=c(2,1))
qqnorm(residuals(reg),pch=16)
qqline(residuals(reg))
boxplot(residuals(reg))
```

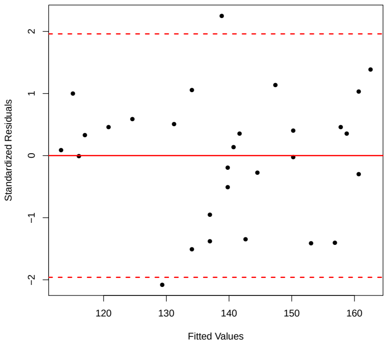


Homogeneity

For a Gaussian LM, the errors are **independant** and **of same variance**
A diagnostic plot consists of looking at the residuals as a function of the fitted values.

```
plot(fitted(reg),rstandard(reg))
abline(h=0,col=2,lwd=2)
abline(h=-1.96,col=2,lwd=2,lty=2)
abline(h=1.96,col=2,lwd=2,lty=2)
```

Blood pressure data

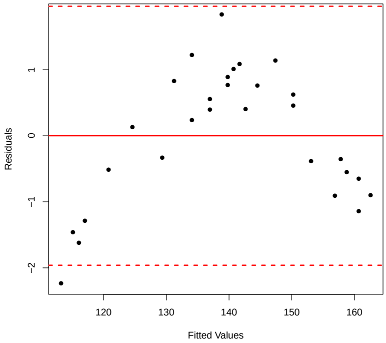


Homogeneity

For a Gaussian LM, the errors are **independant** and **of same variance**
A diagnostic plot consists of looking at the residuals as a function of the fitted values.

```
plot(fitted(reg),rstandard(reg))
abline(h=0,col=2,lwd=2)
abline(h=-1.96,col=2,lwd=2,lty=2)
abline(h=1.96,col=2,lwd=2,lty=2)
```

A pathological case : quadratic term missing



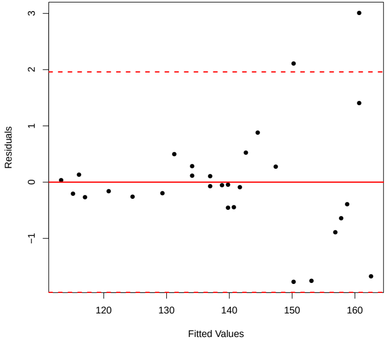
Homogeneity

For a Gaussian LM, the errors are **independant** and **of same variance**

A diagnostic plot consists of looking at the residuals as a function of the fitted values.

```
plot(fitted(reg),rstandard(reg))
abline(h=0,col=2,lwd=2)
abline(h=-1.96,col=2,lwd=2,lty=2)
abline(h=1.96,col=2,lwd=2,lty=2)
```

Heterogeneous variances

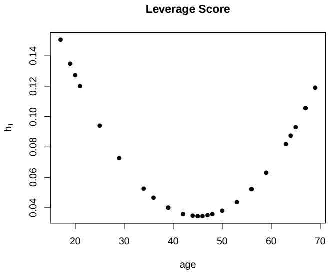


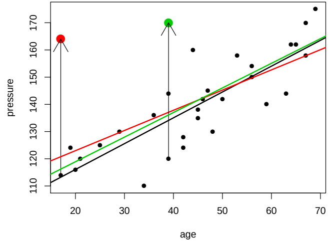
Outliers : leverage

Ina linear regression, isolated points located at the extremities of the observed range of values (for a covariate) can have a strong influence on the estimation of the associated parameter : this is called the **leverage**.

The leverage score of a datapoint *i* is its associated entry in the hat matrix **H** :

$$h_{ii} = (\mathbf{H})_{ii}$$





Outliers : residual analysis

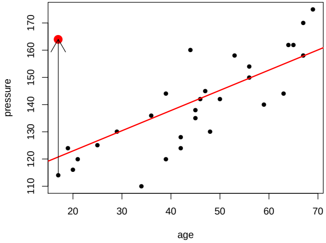
- Residuals are predictions of the errors : $Var(e_i) = \sigma^2$ but $Var(\hat{e}_i) = \sigma^2(1 - h_{ii})$
- Stadardized residuals (R `rstandard`) :

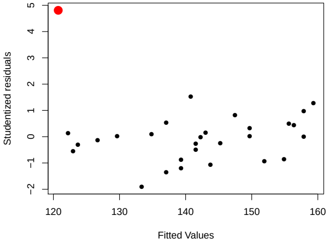
$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}$$

- Studentized residuals (R `rstudent`) :

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{(1 - h_{ii})}} \sim t_{n-3}$$

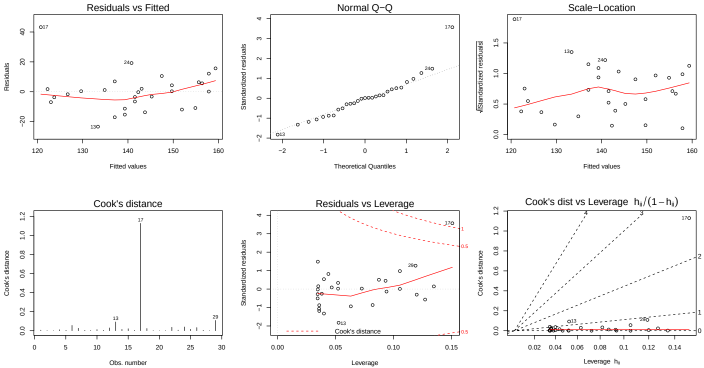
where $\hat{\sigma}_{(i)}$ is an estimation of σ not using the ith data point.
The outlier status of a datapoint can therefore be formally tested $H_0 : t_i \sim t_{n-3}$





Goodness of fit : recap

most of these diagnostics are available in R, using the `plot` funcction on an R object created with the `lm` function (*e.g.* plot(reg,which=1 :6))



Outline

1 Introduction

2 Linear Regression

- Model specification
- Estimation
- Goodness of fit
- Inference on the model parameters

3 ANalysis Of VAriance (ANOVA)

- One factor ANOVA
- 2-way ANOVA

Properties of the estimators

In a normal linear models, the estimators $\hat{\beta}$ are Gaussian and unbiases $(E(\hat{\beta}) = \beta)$. Then :

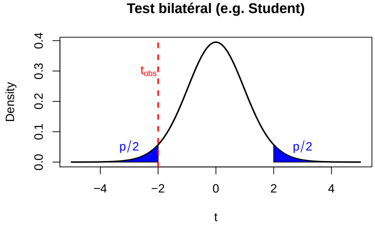
$$\frac{\hat{\beta} - \beta}{\sqrt{\widehat{Var(\hat{\beta})}}} \sim t_{n-2}$$

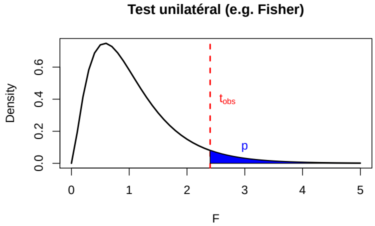
A student's t-distribution with $(n-p)$ degrees of freedom, where p is the number of parameters in the model. This can be used to test the hypothesis $H_0 : \beta = 0$, and construct confidence intervals for β (the parameter) :

$$\left[\hat{\beta} \pm t_{\alpha/2;n-2} \sqrt{(\hat{V}ar(\hat{\beta}))}\right]$$

Hypothesis testing

- Test $H_0 : \beta_2 = 0$ (slope is null) ; or $H_0 : \beta_1 = 0$ (intercept is zero)
- Test statistic $\frac{\hat{\beta}}{\sqrt{(\hat{Var}(\hat{\beta}))}}$ which under H_0 follows a t_{n-2} distribution
- Equivalent to asking whether the confidence interval around β for a given type I error includes 0.
- For categorical variables (ANOVA, we'll get to it in a moment), the test is a compound of multiple hypothesis and the Fisher test is used.


Test bilatéral (e.g. Student)


Test unilatéral (e.g. Fisher)

R output

```
> summary(reg)

Call:
lm(formula = pressure ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-19.354  -4.797   1.254   4.747  21.153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  97.0771     5.5276  17.562 2.67e-16 ***
age           0.9493     0.1161   8.174 8.88e-09 ***
---
Residual standard error: 9.563 on 27 degrees of freedom
Multiple R-squared:  0.7122, Adjusted R-squared:  0.7015
F-statistic: 66.81 on 1 and 27 DF,  p-value: 8.876e-09

> confint(reg)
                 2.5 %     97.5 %
(Intercept) 85.7354850 108.418684
age          0.7110137   1.187631
```

Outline

1 Introduction

2 Linear Regression
- Model specification
- Estimation
- Goodness of fit
- Inference on the model parameters

3 ANalysis Of VAriance (ANOVA)
- One factor ANOVA
- 2-way ANOVA

Classes of Linear Models

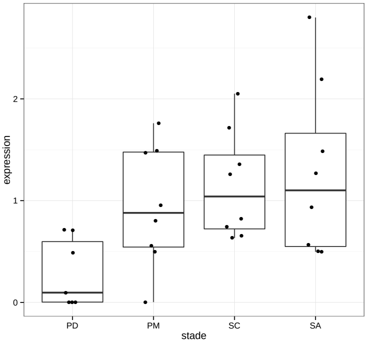
Response	Predictors	Model
Continuous	Binary	t-test
Unique	Categorical	ANOVA
	Continue	Regression (can be multiple)
	Categorical + continue	multiple regression

Outline

1. Introduction
2. Linear Regression
   - Model specification
   - Estimation
   - Goodness of fit
   - Inference on the model parameters
3. ANalysis Of VAriance (ANOVA)
   - One factor ANOVA
   - 2-way ANOVA

Example : expression of gene BMP15 in ovarian cells

- response variable : expression level (RNA transcription)
- Cofactor : cellular developmentalstage (successively PD, PM, SC, SA)
- For each stage i, we have n_i observations $y_{i,1}, y_{i,2}, \dots y_{i,n_i}$

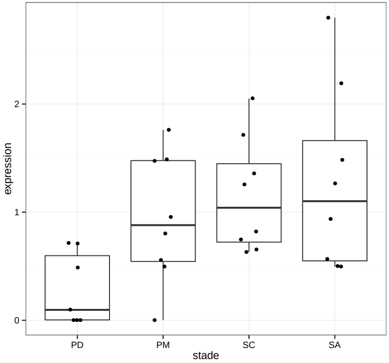


Modèle

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

$j = 1, \ldots n_i$ within-cell index
$i = 1, \ldots p$ cell index (level of the cofactor)
$p = 4$ for BMP15 data
Nomral linear model hypothesis = errors are i.i.d :

$$e_{ij} \sim N(0, \sigma^2)$$



```
> table(bmp15$stade)
PD PM SC SA
 7  8  8  8
```

Link to the multiple regression case

Consider the following model :

$$y_{ij} = \beta_1 x_{ij}^1 + \ldots + \beta_6 x_{ij}^6 + e_{ij}$$

where

- $x_{ij}^s = 1$ if $i = s$ (data is observed at stage s)
- $x_{ij}^s = 0$ otherwise.

It is a multiple regression model, where β_s measures the mean expression level in stage s.

```
  stade expression
1    PD 0.71448438
2    PD 0.70767961
3    PD 0.09622651
4    PM 0.95497190
5    PM 1.48774379
6    PM 1.76053135

...
```

L'ANOVA shares with the multiple regression model

- Estimation procedures (least squares, ML)
- Goodness-of-fit diagnostics

```
  stadePD stadePM stadeSC stadeSA
1       1       0       0       0
2       1       0       0       0
3       1       0       0       0
4       0       1       0       0
5       0       1       0       0
6       0       1       0       0

...
```

Model fitting in R

```
> anov.1=lm(expression~stade,data=bmp15)
> summary(anov.1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2879     0.2320   1.241  0.22532
stadePM       0.6549     0.3177   2.061  0.04902 *
stadeSC       0.8674     0.3177   2.730  0.01101 *
stadeSA       0.9928     0.3177   3.125  0.00422 **

...
```

```
> head(model.matrix(anov.1))
  (Intercept) stadePM stadeSC stadeSA
1           1       0       0       0
2           1       0       0       0
3           1       0       0       0
4           1       1       0       0
5           1       1       0       0
6           1       1       0       0

...
```

Paramétrisation avec une cellule de référence

$$\mu_i = \mu_1 + a_i$$

- μ_1 mean of observations in a reference cell (control level)
- $a_i = \mu_i - \mu_1$ difference between mean of cell i to the reference cell (for $i > 1$)

This is the default parameterization in R.

Another possible parameterization

Une moyenne par niveau : $\mu_1, ..., \mu_p$

```
> anov.2=lm(expression ~ -1 + stade,data=bmp15)
> ## also anov.2=lm(expression ~ 0 + stade,data=bmp15)
> summary(anov.2)
...
Coefficients:
        Estimate Std. Error t value Pr(>|t|)
stadePD   0.2879     0.2320   1.241 0.225317
stadePM   0.9428     0.2170   4.344 0.000177 ***
stadeSC   1.1553     0.2170   5.323 1.28e-05 ***
stadeSA   1.2807     0.2170   5.901 2.75e-06 ***

...
```

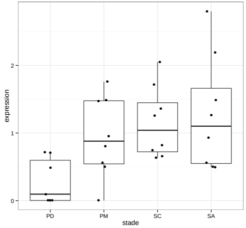
$$y_{ij} = \mu_i + e_{ij}$$

Parameter estimation

- Least squares or maximum likelihood
- $\hat{\mu}_i = \bar{y}_{i\bullet}$ mean of observations in cell i
- $Var(\hat{\mu}_i) = \sigma^2/n_i$: the variance in the estimate of μ is inversely proportional to the number of observations in cell i. ($=$ the *precision* is proportional to n_i).

Testing the cofactor effect : Fisher test

- Involves testing the global hypothesis : $H_0 : \mu_1 = \ldots = \mu_p$ ou $H_0 : a_2 = \ldots = a_p = 0$, all levels have the same mean
- H_1 : ?
- Adding a cofactor effect involves a decomposition of the totale variance $\sum(y_{ij} - \bar{y})^2$ into
  - variance due to differences between level means : "$\sum(a_j - \mu)^2$"
  - Variation within levels : "$\sum(y_{ij} - a_j)^2$"

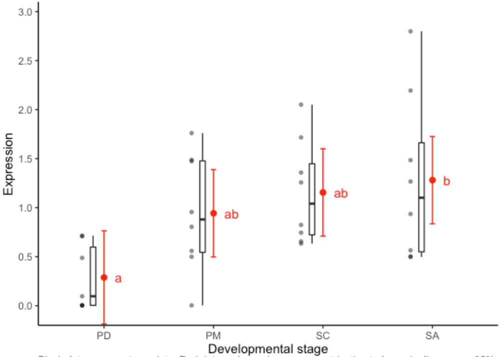


```
> sum((expression-mean(expression))^2)
[1] 14.44973
> anova(anov.1)
Analysis of Variance Table

Response: expression
          Df  Sum Sq Mean Sq F value  Pr(>F)
stade      3  4.2756 1.42519  3.7822 0.02189
Residuals 27 10.1741 0.37682
```

Testing level means : the Tukey HSD test

- Another possible approach : perform all possible pairwise comparisons between factor levels
- Test $\mu_i = \mu_j \forall (i, j)$
- Tukey's Honestly Significant Difference Test
- Corrects for multiple testing
- Usually associated with Compact Letter Display to highlight significant differences



Black dots represent raw data. Red dots and error bars represent (estimated marginal) means ± 95% confidence interval per group. Means not sharing any letter are significantly different by the Tukey-test at the 5% level of significance.

Outline

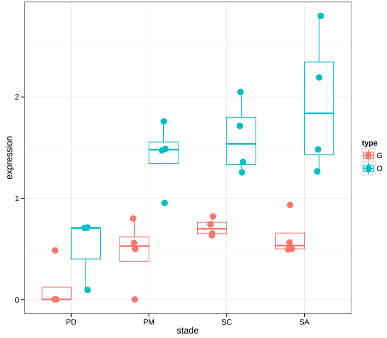
1. Introduction
2. Linear Regression
   - Model specification
   - Estimation
   - Goodness of fit
   - Inference on the model parameters
3. ANalysis Of VAriance (ANOVA)
   - One factor ANOVA
   - 2-way ANOVA

2-way ANOVA

	Facteur 2				
Facteur 1	1	…	j	…	p
1					
…					
i			$\{y_{ijk}\}_{k=1,\dots n_{ij}}$		
…					
q					

n_{ij} observations for levels i of the first factor and j of the second factor (cell)ij).

Example accouting for cell type (**G**ranulosa, **O**ocyte)



Additive Model

$$y_{ijk} = \mu_{ij} + e_{ijk}$$

$$e_{ijk} \sim N(0, \sigma^2) \text{ iid}$$

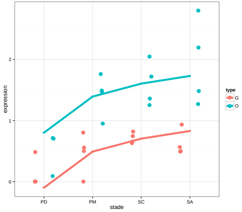
$$i = 1, \dots, p$$

$$j = 1, \dots, q$$

$$k = 1, \dots, n_{ij}$$

			α_1
	μ_{ij}		α_i
			α_p
β_1	β_j	β_q	

$$\mu_{ij} = \alpha_i + \beta_j : p + q \text{ parameters}$$



```
> anova(anov.2)
Analysis of Variance Table

Response: expression
          Df Sum Sq Mean Sq F value    Pr(>F)
stade      3 4.2756  1.4252  9.4494 0.0002144
type       1 6.2527  6.2527 41.4571 8.017e-07
Residuals 26 3.9214  0.1508
```

Interaction Model

$$y_{ijk} = \mu_{ij} + e_{ijk}$$

$$e_{ijk} \sim N(0, \sigma^2) \text{ iid}$$

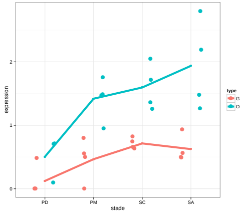
$$i = 1, \ldots, p$$

$$j = 1, \ldots, q$$

$$k = 1, \ldots, n_{ij}$$

			α_1
	μ_{ij}		α_i
			α_p
β_1	β_j	β_q	

$\mu_{ij} = \alpha_i + \beta_j + \gamma_{ij}$: $p \times q$ paramètres



```
> anova(anov.3)
Analysis of Variance Table

Response: expression
           Df Sum Sq Mean Sq F value    Pr(>F)
stade       3 4.2756  1.4252  10.517 0.0001502
type        1 6.2527  6.2527  46.140 6.293e-07
stade:type  3  0.8045   0.2682   1.979  0.14524
Residuals  23 3.1169  0.1355
```

Type I, II and III sums of squares

When the data are unbalanced, there are different ways to compute sums of squares, and therefore test for the effect of a factor :

Type I "Sequential" sums of squares

The results depend on the order the factors appear in the model

- SS(A) for factor A , SS(B|A) for factor B, SS(AB|A,B) for interaction AB
- Default in R

Type I, II and III sums of squares

When the data are unbalanced, there are different ways to compute sums of squares, and therefore test for the effect of a factor :

Type II conditional, no interaction

Tests for each main effect after the other factor

- SS(A|B) for factor A SS(B|A) for factor B
- Assumes no interaction so should only be used if SS(AB|A,B) is not significant
- Default in `python statsmodel`

Type I, II and III sums of squares

When the data are unbalanced, there are different ways to compute sums of squares, and therefore test for the effect of a factor :

Type III, conditional, with interaction

Tests for the presence of a main effect after the other factor and the interaction

- SS(A|B,AB) for factor A, SS(B|A,AB) for factor B
- If interactions are not significant, use type II, otherwise type III.
- Default in `SAS`

Type I, II and III sums of squares

```
> anov.typeI.1 = lm(expression ~stade*type, data=bmp15)
> anova(anov.typeI.1)
           Df Sum Sq Mean Sq F value    Pr(>F)
stade       3 4.2756  1.4252  10.517 0.0001502 ***
type        1 6.2527  6.2527  46.140 6.293e-07 ***
stade:type  3 0.8045  0.2682   1.979 0.1452498
Residuals  23 3.1169  0.1355

> anov.typeI.2 = lm(expression ~type*stade, data=bmp15)
> anova(anov.typeI.2)
           Df Sum Sq Mean Sq F value    Pr(>F)
type        1 6.8262  6.8262 50.3720 3.138e-07 ***
stade       3 3.7021  1.2340  9.1061 0.0003698 ***
type:stade  3 0.8045  0.2682  1.9790 0.1452498

Residuals  23 3.1169  0.1355
```

```
> options(contrasts=c("contr.sum","contr.poly"))
> anov.typeIII = lm(expression ~stade*type, data=bmp15)
> drop1(anov.typeIII,.~., test='F')
Single term deletions

Model:
expression ~ stade * type
           Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                  3.1169 -55.212
stade       3    3.9752 7.0921 -35.725  9.7778 0.0002387 ***
type        1    5.9715 9.0884 -24.037 44.0650 8.996e-07 ***

stade:type  3    0.8045 3.9214 -54.094  1.9790 0.1452498
```