

# Modelos Matemáticos e Aplicações

## The Linear Model

Jorge Cadima

Secção de Matemática (DCEB) - Instituto Superior de Agronomia (UL)

2021-22

## Module 2: Statistical Modelling

Introduction to the main statistical models.

- 1 The Linear Model
- 2 Generalized Linear Models
- 3 Linear Mixed Models

The best-known and most used statistical models are instances of the **Linear Model**.

- Linear Regression (Simple and Multiple)
- Polynomial Regression
- Analysis of Variance (ANOVA)
- Analysis of Covariance (ANCOVA)

# Bibliography - Linear Model

## 1 Notes for the *Estatística e Delineamento* MSc course:

- ▶ Cadima, J. (2020) [O Modelo Linear](#) (in Portuguese only)

## 2 Basic References:

- ▶ Draper, N.R. and Smith, H. (1998), *Applied Regression Analysis*, 3d. edition, John Wiley & Sons **[BISA: U10-734] + [SI-78]** (**[BISA: U10-412]** First Edition, 1981).
- ▶ Kutner, M.H.; Nachtsheim, C.J.; Neter, J. and Li, W. (2005), *Applied Linear Statistical Models*, Irwin **[BISA: U10-727 e CD-236]**.
- ▶ Montgomery, D.C. and Peck, E.A. (1982), *Introduction to Linear Regression Analysis*, John Wiley & Sons **[BISA: U10-329]**.
- ▶ Seber, G.A.F. (1977), *Linear Regression Analysis*, John Wiley & Sons **[BISA: U10-416]**

# Bibliography (continuation)

## 3 References for the use of R

- ▶ Agresti, Alan (2015) *Foundations of Linear and Generalized Linear Models*. Wiley Series in Probability and Statistics.
- ▶ Fox, John and Weisberg, Harvey Sanford (2011) *An R Companion to Applied Regression*, SAGE publications.
- ▶ Maindonald, J. and Brown, W.J. (2003), *Data Analysis and Graphics using R*, Cambridge University Press [**BISA: U10-722**]
- ▶ Venables, W.N. and Ripley, B.D. (2002), *Modern Applied Statistics with S (fourth edition)*, Springer-Verlag [**BISA: U10-733**]

# Statistical Modelling

**Goal:** To study the **relation** between

- a **response variable** (or **dependent variable**)  $y$ ; and
- one or more **predictor variables** (**explanatory** or **independent variables**),  $x_1, x_2, \dots, x_p$ .

This relation is studied based on  $n$  observations of the variables involved in the relation.

# Our models

In this course we only consider models:

- with a single numerical response variable.
- fitted with  $n$  independent observations (does not include time series or spatial data).

As for the predictors:

- there can be one or more predictors;
- the predictors can be numerical or categorical (factors).

We motivate our discussion with some examples.

# Example 1: simple linear regression (descriptive)

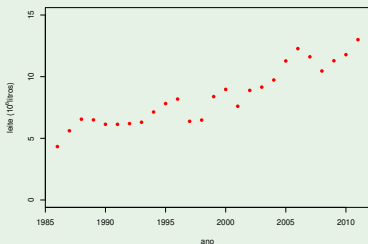
## Example 1: Goat milk

**Response:** Production of goat milk in Portugal ( $y$ , milk) ( $10^6$  litres).

**Predictor:** Years ( $x$ , year) (1986 to 2011).

**Data:**  $n=26$  pairs of values,  $\{(x_i, y_i)\}_{i=1}^{26}$ . In the *data frame* *Cabra*.

**Source:** Portugal's National Statistics Institute (*INE*).



The underlying trend is approximately **linear**.

The focus is on the **descriptive context** (this is not a sample).

What is the “best” equation  $y = b_0 + b_1 x$ , to describe the linear relation with a given set of  $n$  observations (and what does “best” mean)?

## Example 2 - simple linear regression (inferential)

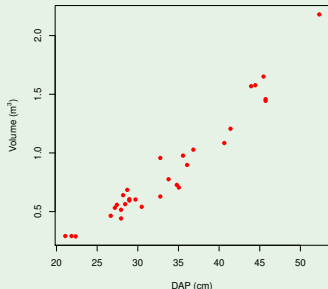
### Example 2: Volume of cherry tree trunks

**Response** (numerical): Volume of the trunks ( $y$ ) of cherry trees.

**Predictor** (numerical): Diameter of the tree trunk at 1.30m. ( $x$ , DAP).

**Data**:  $n=31$  pairs of observations,  $\{(x_i, y_i)\}_{i=1}^{31}$ . *Data frame* `trees`.

**Source**: In R: see `help(trees)` for details. Converted to the metric system.

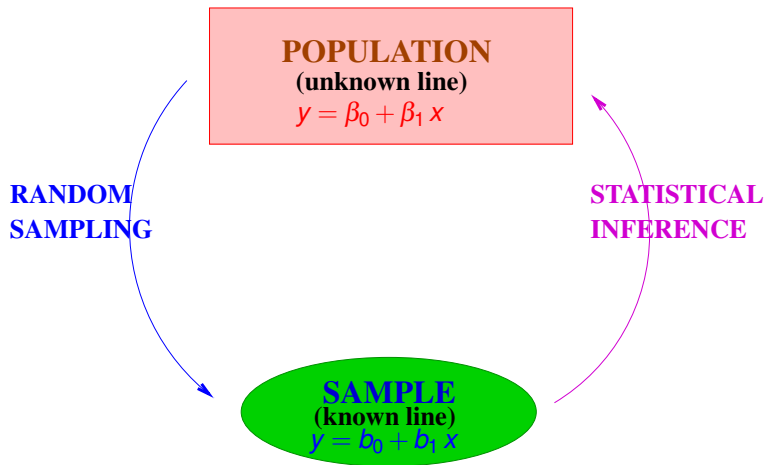


An approximately linear underlying trend.

We have a **random sample** from a much larger populations. We are interested in the **inferential context**: what can we say about the **population** straight line  $y = \beta_0 + \beta_1 x$ ?



# Statistical Inference in a Simple Linear Regression



# Example 3: one-way ANOVA

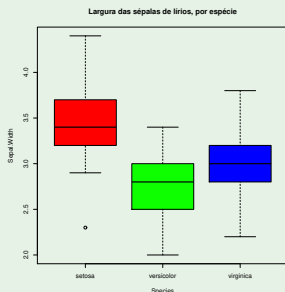
## Example 3: Sepal width in iris

**Response** (numerical): sepal width in *iris* flowers.

**Predictor** (factor): species.

**Data**:  $n = 150$  measurements, 50 for each of 3 species. Data frame `iris`.

**Source**: R: see `help(iris)` for details.



Are there differences in the mean **population** values for each species?

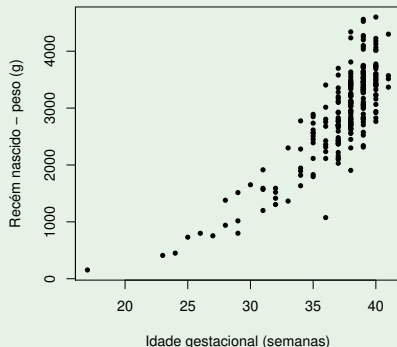
## Example 4 - non-linear relation (descriptive)

### Example 4: Weight of babies at birth

**Response** (numerical): weight of new-born babies ( $y$ ), in g.

**Predictor** (numerical): Duration of pregnancy ( $x$ ), in weeks.

**Data**:  $n = 251$  pairs of observations,  $\{(x_i, y_i)\}_{i=1}^{251}$ .



The underlying trend is clearly **non-linear**:  $y = f(x)$ .

## Example 4 (cont.)

Now, there is a further issue:

- What is the nature of the function  $f$  in  $y = f(x)$ ?
  - ▶  $f$  exponential ( $y = ce^{dx}$ )?
  - ▶  $f$  power law ( $y = cx^d$ )?

Once the class of functions  $f$  is defined, there are similar issues as before: how to determine the “best” parameters  $c$  and  $d$ ?

Non-linear relations are studied by a Non-linear regression (not covered in the MMA course).

But many non-linear relations can be linearised through appropriate transformations of the variables, and the resulting linearised relation can be studied using the Linear Model.

## Example 5 - non-linear relation (inferential)

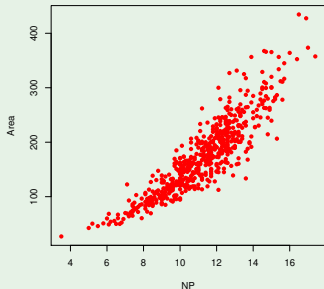
### Example 5: Vineleaves' surface area

**Response** (numerical): Surface area of vine leaves ( $y$ , Area).

**Predictor** (numerical): length of the main vein ( $x$ , NP).

**Data**:  $n = 600$  pairs of observations,  $\{(x_i, y_i)\}_{i=1}^{600}$ . *Data frame* `videiras`.

**Source**: Prof. Carlos Lopes, Viticulture, ISA.



**Non-linear** trend  $y = f(x)$ . Parabolic? Exponential? Power function?

Data are a **random sample**. What can be said about the parameters in the **population**?

## Example 6 - ANCOVA-type relation

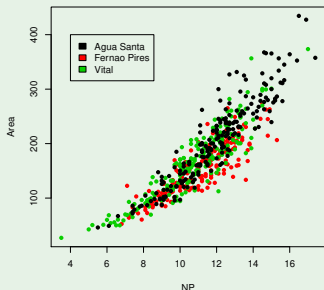
### Example 6: Vineleaves' surface area

**Response** (numerical): Surface area of vine leaves ( $y$ , Area).

**Predictor** (numerical): length of the main vein ( $x$ , NP).

**Predictor** (factor): variety (3 varieties: Água Santa, Fernão Pires and Vital).

**Data**:  $n = 200$  observations for each variety. *Data frame* `videiras`.



Does a single curve fit all varieties well?

Or are different curves for different varieties preferable?

# Example 7 - Multiple linear regression

## Example 7: Anthocyanine content

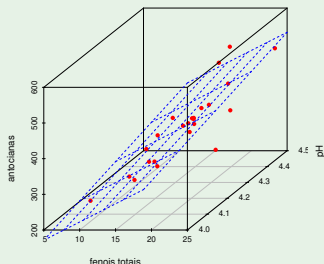
**Response** (numerical): Anthocyanine content ( $y$ ,  $\text{antoci}$ ) (in  $\text{mg}/\text{dm}^3$ ).

**Predictor** (numerical): Total phenol content ( $x_1$ ,  $\text{fentot}$ ).

**Predictor** (numerical): pH ( $x_2$ ,  $\text{pH}$ ).

**Data**:  $n=24$  genotypes of the Tinta Francisca variety. *Data frame*  $\text{Antoci}$ .

**Source**: Prof. Elsa Gonçalves, ISA (Tabuaço 2003).



**Descriptive**: what is the “best” **sample plane**  $y = b_0 + b_1x_1 + b_2x_2$ ?

**Inferential**: what can be said about the **population plane**  $y = \beta_0 + \beta_1x_1 + \beta_2x_2$ ?

# Modelling: initial considerations

- All models are just **approximations** of reality.
- There may be **different suitable models**.
- The **principle of parsimony**: among models considered **suitable**, simpler ones are to be preferred.
- Models may be:
  - ▶ **theoretical**, based on physical, biological or other principles;
  - ▶ **empirical**, describing a relation observed in the data.
- **Statistical** models are not deterministic: they describe **underlying trends**, but there is **variability** around that trend. This variability should be **incorporated into the model**.



## Initial considerations (cont.)

- There need not be a cause-and-effect relation between predictors and response variables. Statistics deals with association. A possible cause-and-effect relation can only be shown by extra-statistical considerations.
- Different approaches may exist when studying statistical models:
  - ▶ **descriptive**: fitting a model to highlight relations in the data, regardless of their origin.
  - ▶ **inferential**: when the data are a random sample from a population, we seek to draw conclusions regarding the population.

Inference requires more assumptions and a much heavier mathematical-statistical framework.

# The Linear Model

- The **Linear Model** is **one category** of statistical models;
- it encompasses many different more specific models:  
Linear Regressions (Simple and Multiple), Polynomial Regression,  
Analysis of Variance, Analysis of Covariance;
- is the **most used type of model**, with a long tradition;
- it serves as a **reference for numerous generalizations**:  
Nonlinear Regression; Generalized Linear Models;  
Mixed Linear Models, etc.

# Review: descriptive Simple Linear Regression

Given  $n$  pairs of observations  $\{(x_i, y_i)\}_{i=1}^n$ , we have:

The regression line of  $y$  over  $x$

$$y = b_0 + b_1 x$$

is given by:

$$\text{Slope } b_1 = \frac{\text{COV}_{xy}}{s_x^2} \quad \left( \frac{\text{units of } y}{\text{units of } x} \right)$$

$$\text{Intercept } b_0 = \bar{y} - b_1 \bar{x} \quad (\text{units of } y)$$

with

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i & ; & \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \\ s_x^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} & ; & \quad \text{cov}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{n-1} . \end{aligned}$$

## Review: descriptive Simple Linear Regression (cont.)

How was this equation obtained?

**Criterion:** Minimize the sum of squared residuals (Legendre 1805, Gauss 1795-1809).

### Residuals and Residual Sum of Squares

**Residuals** are (signed, vertical) distances between each point and the line:

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i),$$

where  $\hat{y}_i = b_0 + b_1 x_i$  are the “values of  $y$ , fitted by the regression line”.

**Residual Sum of Squares (RSS):**

$$SQRE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2.$$

**Criterion:** Determine the  $b_0$  and  $b_1$  that minimise  $SQRE$ .

**Note:**  $SQRE$  has units of measurement: the square of the units of  $y$ .

## Review: Descriptive Simple Linear Regression (cont.)

To minimise  $SQRE$ , its partial derivatives with respect to  $b_0$  and  $b_1$  must be set to zero:

$$SQRE(b_0, b_1) = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2$$

$$\begin{cases} \frac{\partial SQRE}{\partial b_0}(b_0, b_1) = 0 \\ \frac{\partial SQRE}{\partial b_1}(b_0, b_1) = 0 \end{cases} \Leftrightarrow \begin{cases} (-2) \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)] = 0 \\ 2 \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)] (x_i) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^n y_i - n b_0 - b_1 \sum_{i=1}^n x_i = 0 \\ \sum_{i=1}^n y_i x_i - b_0 \sum_{i=1}^n x_i - b_1 \sum_{i=1}^n x_i^2 = 0 \end{cases} \Leftrightarrow \begin{cases} b_0 = \bar{y} - b_1 \bar{x} \\ b_1 = \frac{cov_{xy}}{s_x^2} \end{cases}$$

This **critical point** is a **minimum**, because function  $SQRE$  is quadratic and always positive.

# Descriptive Simple Linear Regression with R

Linear regressions are fitted in R using the command `lm` (the initials of **l**inear **m**odel).

The command `lm` has two main arguments:

- `formula` – identifies the **response variable** and the **predictors**; in a simple linear regression of variable  $y$  over the predictor  $x$ :  $y \sim x$ .
- `data` – indicates the name of the *data frame* containing the data.

## R command for the linear regression in Example 1

```
> lm( leite ~ ano , data=Cabra )
```

```
Call: lm(formula = leite ~ ano, data = Cabra)
```

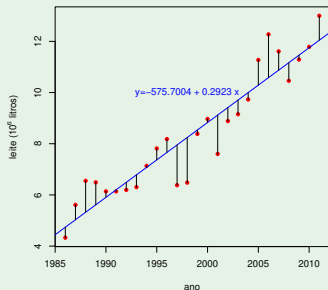
```
Coefficients:
```

```
(Intercept)      ano  
-575.7004      0.2923  <- fitted values of b0 and b1
```

# Descriptive Simple Linear Regression - Example 1

## Example 1: Goat Milk

$x$  - Year ;  $y$  - goat milk production ;  $n=26$  pairs  $\{(x_i, y_i)\}_{i=1}^{26}$ .



The fitted line **minimises the sum of squared vertical distances** between points and line.

# The parameters of the regression line

## Properties of the parameters

- The intercept  $b_0$ :
  - ▶ is the **value of  $y$  (on the line) corresponding to  $x = 0$** ;
  - ▶ has units of measurement equal to those of  $y$ .
- The slope  $b_1$ :
  - ▶ is the **(mean) difference in  $y$  corresponding to an increase of one unit in  $x$** ;
  - ▶ has units of measurement equal to  $\frac{\text{units of } y}{\text{units of } x}$ .

## Example 1: Goat milk

The fitted slope  $b_1 = 0.2923$  means that, on average, the production of goat milk increased  $0.2923 \times 10^6$  litres per year.



# Additional properties of the regression line

## Properties of the regression line

- The regression line always crosses the centre of gravity of the scatterplot, that is, point  $(\bar{x}, \bar{y})$ .

Given the formula for the intercept:  $b_0 = \bar{y} - b_1 \bar{x} \Leftrightarrow \bar{y} = b_0 + b_1 \bar{x}$ .

- The mean of the observed values  $y_i$  equals the mean of the fitted values  $\hat{y}_i$ :  $\bar{y} = \bar{\hat{y}}$ .

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n (b_0 + b_1 x_i) = \frac{1}{n} \underbrace{\sum_{i=1}^n b_0}_{=nb_0} + b_1 \frac{1}{n} \underbrace{\sum_{i=1}^n x_i}_{=\bar{x}} = b_0 + b_1 \bar{x} = \bar{y}.$$

- The mean (and sum) of residuals is zero:  $\bar{e} = 0$ .

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = \frac{1}{n} \underbrace{\sum_{i=1}^n y_i}_{=\bar{y}} - \frac{1}{n} \underbrace{\sum_{i=1}^n \hat{y}_i}_{=\bar{\hat{y}}} = \bar{y} - \bar{\hat{y}} = 0.$$

## R commands to study a regression

Save the regression for Example 1:

```
> Cabra.lm <- lm( leite ~ ano , data=Cabra )
```

- `fitted` gives the fitted values  $\hat{y}_i = b_0 + b_1 x_i$ :

```
> fitted(Cabra.lm)
```

1	2	3	4	5	6	7	8	9	10
4.737154	5.029418	5.321683	5.613948	5.906212	6.198477	6.490742	6.783006	7.075271	7.367535
11	12	13	14	15	16	17	18	19	20
7.659800	7.952065	8.244329	8.536594	8.828858	9.121123	9.413388	9.705652	9.997917	10.290182
21	22	23	24	25	26				
10.582446	10.874711	11.166975	11.459240	11.751505	12.043769				

## R commands (cont.)

- `residuals` gives the residuals  $e_i = y_i - \hat{y}_i$ :

```
> residuals(Cabra.lm)
```

```
      1      2      3      4      5      6      7      8  
-0.40915385  0.58058154  1.22831692  0.87805231  0.23178769 -0.06247692 -0.29474154 -0.47900615  
      9     10     11     12     13     14     15     16  
 0.05772923  0.44946462  0.52220000 -1.57206462 -1.76532923 -0.15359385  0.13814154 -1.52012308  
     17     18     19     20     21     22     23     24  
-0.52738769 -0.55265231 -0.26891692  0.98281846  1.69155385  0.73428923 -0.70797538 -0.17124000  
     25     26  
 0.03249538  0.95723077
```

The Residual Sum of Squares, *SQRE*, can be obtained as follows:

```
> sum(residuals(Cabra.lm)^2)
```

```
[1] 18.04768
```

*SQRE* has units of measurement: the squared units of  $y$ .

## R commands for regression (cont.)

- `predict` – predicts fitted values of new observations given in a *data frame* (the name of the predictor must be the same as in the fitted regression).

```
> novos <- data.frame( ano=c(1985, 2012) )  
> predict( Cabra.lm , new=novos )
```

```
      1      2  
4.444889 12.336034
```

The value  $\hat{y}$  fitted by the regression line, for  $x=2012$ , is:

$$\hat{y} = b_0 + b_1 x$$
$$\Leftrightarrow 12.336034 = -575.7004 + 0.2923 \times 2012 .$$

# The Least Squares criterion

The criterion of minimising the Residual Sum of Squares  $SQRE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$  assumes that:

In a simple linear regression, the role of the variables  $x$  and  $y$ , is **not** symmetric.

**the response variable  $y$**  is the **variable we wish to model**, using variable  $x$ .

**the predictor  $x$**  is the **variable we assume known**, used to draw conclusions regarding  $y$ .

The  $y$  over  $x$  regression line is different from the  $x$  over  $y$  regression line.

## The Least Squares criterion (cont.)

The  $i$ -th residual is the (signed) deviation of observation  $y_i$  in relation to the corresponding value predicted by the regression line:

$$e_i = y_i - \hat{y}_i$$

Minimising the sum of squared residuals means minimising the sum of squared “prediction errors”.

The underlying concern for the criterion is predict variable  $y$  as well as possible, based on its relation with predictor  $x$ .

## Review: The three Sums of Squares

Recall:  $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$  is the sample variance of observations  $y_i$ .

### Total Sum of Squares (SQT)

$$\text{Total SS (SQT)} \quad \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1) s_y^2$$

We have:  $s_{\hat{y}}^2 = \frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  is the sample variance of the fitted  $\hat{y}_i$ .

### Regression Sum of Squares (SQR)

$$\text{Regression SS (SQR)} \quad \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (n-1) s_{\hat{y}}^2$$

### Residual Sum of Squares (SQRE) - already considered

$$\text{Residual SS (SQRE)} \quad \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = (n-1) s_e^2$$

## Review: descriptive simple linear regression (cont.)

### Fundamental Formula of Regression

$$SQT = SQR + SQRE \quad \Leftrightarrow \quad s_y^2 = s_{\hat{y}}^2 + s_e^2$$

### Coefficient of Determination

$$R^2 = \frac{SQR}{SQT} = \frac{s_{\hat{y}}^2}{s_y^2} \in [0, 1]$$

$R^2$  gives the proportion of the total variability of the response variable  $Y$  that is accounted for by the regression. The larger, the better.



# Properties of the Coefficient of Determination

## Properties of $R^2 = \frac{SQR}{SQT}$

- $0 \leq R^2 \leq 1$  (All the SSs are non-negative and  $SQT = SQR + SQRE$ )
- $R^2 = 1$  if, and only if, all  $n$  points are collinear. (“ideal”)  
( $SQT = SQR \Leftrightarrow SQRE = 0$ . Therefore, all residuals are zero: the points are all on the line.)
- $R^2 = 0$  if, and only if, the regression line is horizontal. (“useless”)  
( $SQR = 0 \Leftrightarrow SQRE = SQT$ . All variability of  $y$  is residual, there is no variability among the  $\hat{y}_i$ s (they are all the same). The regression line is  $y = \bar{y} \Leftrightarrow b_1 = 0$ )
- In a **simple** linear regression,  $R^2$  is the squared coefficient of linear correlation between  $x$  and  $y$  (See Exercises):

$$R^2 = r_{xy}^2 = \left( \frac{COV_{xy}}{s_x s_y} \right)^2 \quad \text{if } s_x \neq 0 \text{ and } s_y \neq 0$$

## Example 1: goat milk

The coefficient of determination  $R^2$  is obtained using the command `summary` on a fitted regression. The output says `Multiple R-Squared`.

```
> summary(Cabra.lm)
```

```
Call: lm(formula = leite ~ ano, data = Cabra)
```

```
[...]
```

```
Residual standard error: 0.8672 on 24 degrees of freedom
```

```
Multiple R-squared: 0.8738, Adjusted R-squared: 0.8685
```

```
F-statistic: 166.1 on 1 and 24 DF, p-value: 2.807e-12
```

The value of  $R^2$  (with greater precision) can be obtained as follows:

```
> summary(Cabra.lm)$r.sq
```

```
[1] 0.8737681
```

# Extracting information from a fitted regression

The `lm` command creates an object of type `list`:

```
> is.list(Cabra.lm) <- asks whether Cabra.lm is a list
```

```
[1] TRUE
```

```
> names(Cabra.lm) <- requests the names of the list components
```

```
"coefficients" "residuals" "effects" "rank" "fitted.values" "assign"  
"qr" "df.residual" "xlevels" "call" "terms" "model"
```

Each list component can be extracted by writing the list and component names, separated by a dollar sign:

```
> Cabra.lm$coef <- component name may be incomplete if unambiguous
```

```
(Intercept)          ano  
-575.7003723    0.2922646
```

For more information on each list element: `help(lm)`.

## Extracting information from a regression (cont.)

The `summary` command, applied on a fitted regression, produces a `second object` of type `list`. Here are its components:

```
> names(summary(Cabra.lm))
```

```
[1] "call"           "terms"          "residuals"     "coefficients"  
[5] "aliased"        "sigma"          "df"            "r.squared"  
[9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

Individual components may be extracted from this output list, as seen before: `summary(Cabra.lm)$r.sq` gives the value of  $R^2$ .

## Regression - a bit of History

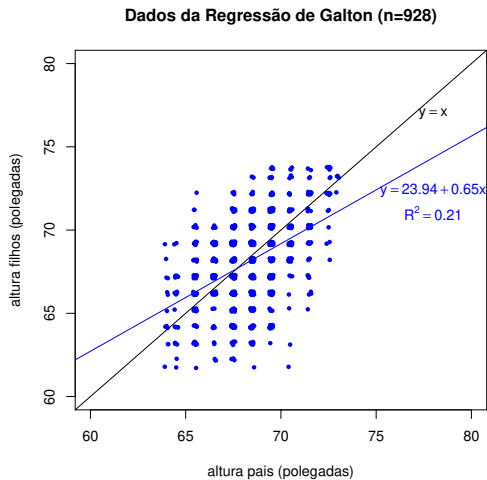
The name **Regression** has its origins in a study by Francis Galton (1886), relating the height of  $n = 928$  young adults with the (mean) height of their parents. Galton invented the term **eugenics**, a concept that was considered reputable until the early 20th century.

Galton noted that parents with above average heights tended to have children with heights above the mean - but less tall than their parents (likewise for those below the mean height).

Galton called his article *Regression towards mediocrity in hereditary stature*. This is the origin of the association of the name **regression** with the method.

Curiously, Galton's dataset has a very low Coefficient of Determination.

# A bit of History (cont.)



# A disadvantage of the Least Squares criterion

The fitting criterion (minimise  $SQRE$ ) is sensitive to outliers.

We illustrate with a dataset from R's `MASS` package (initials of the book *Modern Applied Statistics with S*, by Venables and Ripley).

## Animals - MASS package

```
> library(MASS)    ← to load package MASS
> help(Animals)
```

```
Animals                package:MASS                R Documentation
[...]
Average brain and body weights for 28 species of land animals.
[...]
'body' body weight in kg.
'brain' brain weight in g.
[...]
Source:
P. J. Rousseeuw and A. M. Leroy (1987) _Robust Regression and
Outlier Detection. Wiley, p. 57.
```

# Example: Animals dataset

## > Animals

	body	brain
Mountain beaver	1.350	8.1
Cow	465.000	423.0
Grey wolf	36.330	119.5
Goat	27.660	115.0
Guinea pig	1.040	5.5
Dipliodocus	11700.000	50.0
Asian elephant	2547.000	4603.0
Donkey	187.100	419.0
Horse	521.000	655.0
Potar monkey	10.000	115.0
Cat	3.300	25.6
Giraffe	529.000	680.0
Gorilla	207.000	406.0
Human	62.000	1320.0
African elephant	6654.000	5712.0
Triceratops	9400.000	70.0
Rhesus monkey	6.800	179.0
Kangaroo	35.000	56.0
Golden hamster	0.120	1.0
Mouse	0.023	0.4
Rabbit	2.500	12.1
Sheep	55.500	175.0
Jaguar	100.000	157.0
Chimpanzee	52.160	440.0
Rat	0.280	1.9
Brachiosaurus	87000.000	154.5
Mole	0.122	3.0
Pig	192.000	180.0

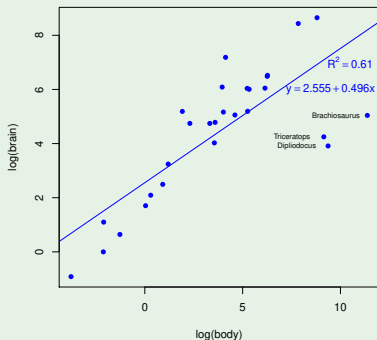


# Simple linear regression and outliers

## Example: Animals

Most observations follow a **linear relation** between the **logarithms** of brain and body weights.

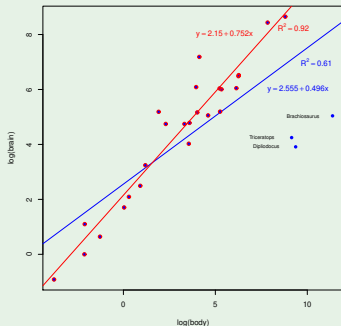
But three species of dinosaurs are **outliers** and affect the fitted line.



# Simple linear regression and outliers (cont.)

## Exemplo: Animals

Excluding those observations changes the fitted line and its quality.



In this case, we can exclude the 3 outliers because they are from a “different reality” (extinct species). There are **alternative fitting criteria** that are **robust**.

# Non-linear relations and linearizing transformations

In some cases, an underlying non-linear trend between  $x$  and  $y$  can be linearized by suitable transformations of one, or both, variables.

Such transformations enable us to apply simple linear regressions even when the original relation is non-linear.

These linearizing transformations can also be useful when there is more than one predictor.

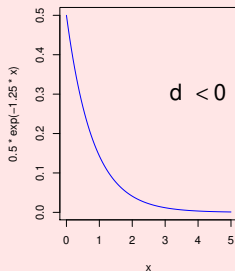
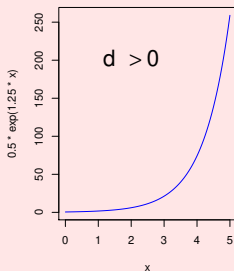
We consider some particularly frequent examples of non-linear relations that can be linearized by transformations of the response variable and, in some cases, also of the predictor.

# The exponential relation

## Exponential relation

$$y = ce^{dx}$$

$$(y > 0 ; c > 0)$$



Linearizing transformation:  $y^* = \ln(y)$  and  $x^* = x$

# Linearizing an exponential relation

Taking **logarithms** in the exponential equation, we obtain a **linear relation** between  $y^* = \ln(y)$  and  $x$ :

$$\begin{aligned}y = ce^{dx} &\Leftrightarrow \ln(y) = \ln(c) + \ln(e^{dx}) = \ln(c) + dx \\ &\Leftrightarrow y^* = b_0 + b_1 x\end{aligned}$$

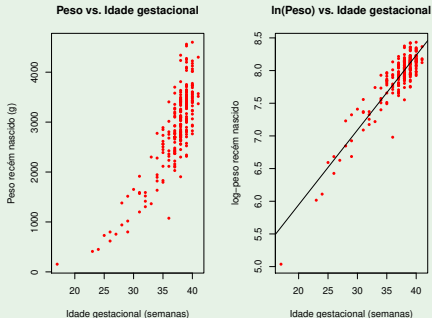
with **slope**  $b_1 = d$  and **intercept**  $b_0 = \ln(c)$ .

The **sign of the line's slope** indicates whether the original exponential relation is **increasing** ( $b_1 > 0$ ) or **decreasing** ( $b_1 < 0$ ).

# Linearizing the relation

## Example 4: weight of babies at birth

A scatterplot of **log-weights of new-born babies** versus the duration of pregnancy shows an **underlying linear relation**:



This linearization means that the **original relation (weight vs. duration of pregnancy)** may be considered **exponential**.

# The exponential relation

## Differential equation corresponding to the exponential

An exponential relation results from assuming that  $y$  is a function of  $x$  and that the **rate of change of  $y$** , that is, the derivative  $y'(x)$ , is proportional to  $y$ :

$$y'(x) = d \cdot y(x) ,$$

i.e., that the **relative rate of change of  $y$**  is constant:

$$\frac{y'(x)}{y(x)} = d .$$

Integrating (in relation to  $x$ ), we have (since  $y > 0$ ):

$$\ln|y(x)| = dx + K \quad \Leftrightarrow \quad y(x) = e^{K+dx} \quad \Leftrightarrow \quad y(x) = e^K e^{dx} .$$

The slope of the line  $b_1$  is the constant  $d$  relative rate of change of  $y$ .

The integration constant  $K$  is the intercept:  $K = b_0$ .

# Logistic model for population growth

An exponential model is frequently used to describe **population growth**, in an initial phase where the impact of limiting resources is not yet felt. But **exponential population growth is not sustainable in the long run**.

In 1838 Verhulst<sup>1</sup> suggested an **alternative model for population growth**, which accounted for effects of resource shortages: the **logistic model**.

We consider here a **simplified (2 parameter) version** of a logistic curve, associated with a response variable that gives the **proportion of the carrying capacity of the environment** (size of population in relation to its maximum possible value).

---

<sup>1</sup>Verhulst, P.-F. (1838), Notice sur la loi que la population poursuit dans son accroissement. *Corresp. Math. Phys.* **10**, 113-121

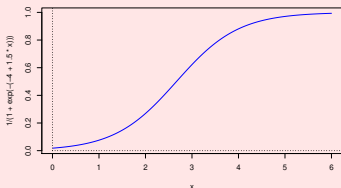


# Logistic relation (with 2 parameters)

## Two-parameter logistic relation

$$y = \frac{1}{1 + e^{-(c+dx)}}$$

$$(y \in ]0, 1[)$$



( $d > 0$ )

Linearizing transformation: *logit* transformation of  $y$ , i.e.,

$$y^* = \ln\left(\frac{y}{1-y}\right) \quad \text{e} \quad x^* = x$$

# Linearizing the logistic relation

Since  $y \in ]0, 1[$ , the *logit transformation*,  $y^* = \ln\left(\frac{y}{1-y}\right)$ , is well defined.

The logistic relation between  $y$  and  $x$  corresponds to a **linear relation** between  $y^* = \ln\left(\frac{y}{1-y}\right)$  and  $x^* = x$ :

$$\begin{aligned}y &= \frac{1}{1 + e^{-(c+dx)}} &\Leftrightarrow & 1 - y = 1 - \frac{1}{1 + e^{-(c+dx)}} = \frac{e^{-(c+dx)}}{1 + e^{-(c+dx)}} \\& &\Leftrightarrow & \frac{y}{1-y} = \frac{1}{e^{-(c+dx)}} = e^{c+dx} \\& &\Leftrightarrow & \underbrace{\ln\left(\frac{y}{1-y}\right)}_{=y^*} = \underbrace{c}_{=b_0} + \underbrace{d}_{=b_1} x\end{aligned}$$

## More on the Logistic

### Differential equation of the logistic (with 2 parameters)

The logistic relation results from assuming that  $y$  is a function of  $x$  and that  $y$ 's relative rate of growth decreases linearly with the growth of  $y$ :

$$\frac{y'(x)}{y(x)} = d \cdot [1 - y(x)] .$$

The previous equation is equivalent to:

$$\frac{y'(x)}{y(x) \cdot [1 - y(x)]} = d \quad \Leftrightarrow \quad \frac{y'(x)}{y(x)} + \frac{y'(x)}{1 - y(x)} = d$$

Integrating (in relation to  $x$ ), gives (since  $\int \frac{f'}{f} = \ln(|f|)$ ):

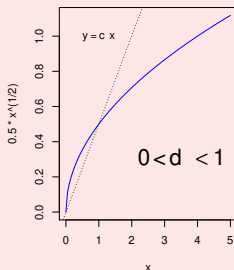
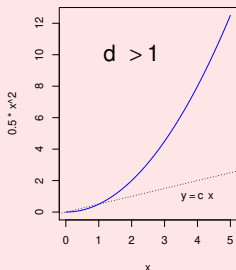
$$\begin{aligned} \ln y(x) - \ln(1 - y(x)) &= dx + K \\ \Leftrightarrow \ln\left(\frac{y}{1 - y}\right) &= b_1 x + b_0 . \end{aligned}$$

# Power (or allometric) relation

## Power law

$$y = cX^d$$

$$(x, y > 0 \ ; \ c > 0)$$



Linearizing transformation:  $y^* = \ln(y)$  and  $x^* = \ln(x)$ .

# The linearization of a power relation

Taking logarithms, we have:

$$\begin{aligned}y = cX^d &\Leftrightarrow \ln(y) = \ln(cX^d) = \ln(c) + \ln(X^d) \\&\Leftrightarrow \ln(y) = \ln(c) + d \ln(X) \\&\Leftrightarrow y^* = b_0 + b_1 x^*\end{aligned}$$

which is a **linear relation between  $y^* = \ln(y)$  and  $x^* = \ln(x)$** .

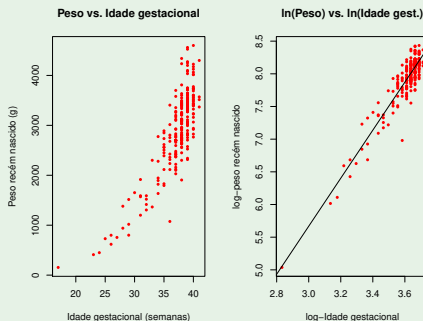
The slope  $b_1$  of the line is the exponent  $d$  in the power law.

The intercept is  $b_0 = \ln(c)$ , that is,  $c = e^{b_0}$ .

# Another linearization in Example 4

## A different linearization for the weight of babies

The scatterplot of **log-weights** of new-born babies vs. the **log-duration of pregnancy** results in another underlying linear trend:



This linearization means that the **original relation** (weight vs. duration of pregnancy) can **also** be considered a power relation.

# More on the power relation

## A differential equation for a power relation

A power relation results from assuming that  $y$  is a function of  $x$  and the **relative rate of growth of  $y$** , i.e., the ratio  $\frac{y'(x)}{y(x)}$ , is inversely proportional to  $x$ :

$$\frac{y'(x)}{y(x)} = \frac{d}{x}.$$

Integrating (in relation to  $x$ ), gives (since  $y > 0$  and  $x > 0$ ):

$$\underbrace{\ln|y(x)|}_{=y^*} = \underbrace{d}_{=b_1} \underbrace{\ln|x|}_{=x^*} + \underbrace{K}_{=b_0} \quad \Leftrightarrow \quad y(x) = e^{K+\ln(x^d)} \quad \Leftrightarrow \quad y(x) = e^K x^d.$$

The line's slope  $b_1$  is the constant of (inverse) proportionality  $d$ .

The constant of integration  $K$  is the line's intercept:  $K = b_0$ .

# Another differential equation for the power relation

## The allometric differential equation

A different way of obtaining a power relation, used in the study of allometry, is to assume that  $y$  and  $x$  are both functions of a third variable  $t$  (i.e.,  $y(t)$  and  $x(t)$ ) and that the relative rates of growth of  $y$  and  $x$  are proportional:

$$\frac{y'(t)}{y(t)} = d \cdot \frac{x'(t)}{x(t)}.$$

Integrating (in relation to  $t$ ) gives:

$$\ln y = d \ln x + K$$

and exponentiating,

$$y = e^{d \ln x + K} = e^{d \ln x} \cdot e^K = x^d \cdot \underbrace{e^K}_{=c} \Leftrightarrow y = cx^d.$$

Studies of **allometry** compare the size of different parts of an organism. **Isometry** results when  $d=1$ .

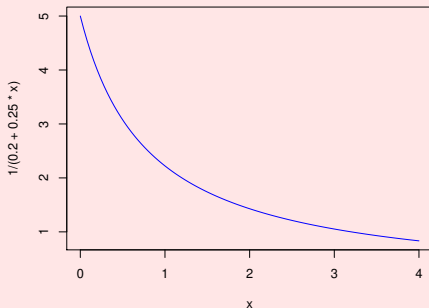


# A hyperbolic (or inverse proportionality) relation

## Hyperbolic-type relation

$$y = \frac{1}{c + dx}$$

$$(x, y > 0 \quad ; \quad c, d > 0)$$



Linearizing transformation:  $y^* = 1/y$  and  $x^* = x$

# The linearization of a hyperbolic relation

Taking **reciprocals** in a hyperbolic-type relation, gives a **linear relation between  $y^* = \frac{1}{y}$  and  $x$** :

$$\begin{aligned}y &= \frac{1}{c + dx} && \Leftrightarrow && \frac{1}{y} = c + dx \\ & && \Leftrightarrow && y^* = b_0 + b_1 x.\end{aligned}$$

with  $b_0 = c$  and  $b_1 = d$ .

Relations of a hyperbolic type have been used, in Agronomy, to model the relation between **yield per plant ( $y$ )** and **crop density ( $x$ )**, for some crops.

**Attention:** For values of  $y$  close to zero, the reciprocal becomes very large. Observations with  $y_i \approx 0$  tend to dominate the fit in a linearized relation.

# More about hyperbolic-type relations

## Differential equation for a hyperbolic-type relation

Assume that the (decrease) in the rate of variation of  $y$  is proportional to the square of  $y$ :

$$y'(x) = -d y^2(x)$$

or equivalently, that the relative rate of growth of  $y$  is proportional to  $y$ :

$$\Leftrightarrow \frac{y'(x)}{y(x)} = -d y(x).$$

Re-writing the equation as  $\frac{y'(x)}{y^2(x)} = -d$ , and integrating  $\left(\int f^\alpha \cdot f' = \frac{f^{\alpha+1}}{\alpha+1}\right)$ , we have:

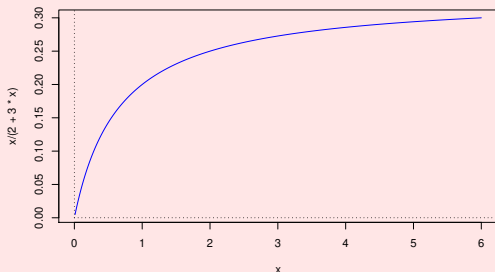
$$-\underbrace{\frac{1}{y(x)}}_{=y^*} = -\underbrace{d}_{=b_1} x + \underbrace{K}_{=b_0} \Leftrightarrow y(x) = \frac{1}{d x + c},$$

with  $c = -K$ . The constant of proportionality ( $-d$ ) is minus the slope of the line ( $b_1$ ).

# Michaelis-Menten relation

## Michaelis-Menten relation

$$y = \frac{x}{c + dx}$$



The horizontal line  $y = \frac{1}{d}$  is an asymptote on the right.

Linearizing transformation:  $y^* = \frac{1}{y}$  e  $x^* = \frac{1}{x}$

## Linearizing the Michaelis-Menten relation

Taking reciprocals in the Michaelis-Menten relation, we obtain a linear relation between  $y^* = \frac{1}{y}$  and  $x^* = \frac{1}{x}$ :

$$\begin{aligned}y = \frac{x}{c + dx} &\Leftrightarrow \frac{1}{y} = \frac{c + dx}{x} \\ &\Leftrightarrow \frac{1}{y} = \frac{c}{x} + d = c \cdot \frac{1}{x} + d \\ &\Leftrightarrow y^* = b_0 + b_1 x^*,\end{aligned}$$

with  $b_0 = d$  e  $b_1 = c$ .

**Attention:** For values of  $y$  or  $x$  close to zero, the reciprocals become very large. Observations with  $y_i \approx 0$  and/or  $x_i \approx 0$  tend to dominate the fit in the linearized relation.

## Michaelis-Menten relation (cont.)

- The Michaelis-Menten relation is used in the study of **enzymatic reactions**, relating the **rate of reaction** with the **concentration of the substrate**.
- In **agronomical yield models** it is known as the **Shinozaki-Kira** model, with  $y$  giving the **total yield** and  $x$  the crop **density**.
- In **fisheries** it is known as the **Beverton-Holt** model:  $y$  is the **recruitment** (size of the next generation) and  $x$  is the size of the **stock** (previous generation).

## Michaelis-Menten relation (cont.)

### Differential equation for a Michaelis-Menten relation

A Michaelis-Menten relation results by assuming that  $y$  is a function of  $x$  and the growth rate of  $y$  is proportional to the squared ratio of  $y$  over  $x$ :

$$y'(x) = c \left[ \frac{y(x)}{x} \right]^2 .$$

Re-writing the equation as  $\frac{y'(x)}{y^2(x)} = c \frac{1}{x^2}$ , and integrating  $\left( \int f^\alpha \cdot f' = \frac{f^{\alpha+1}}{\alpha+1} \right)$ , we have:

$$\begin{aligned} -\frac{1}{y(x)} &= -c \frac{1}{x} + K \Leftrightarrow \underbrace{\frac{1}{y(x)} = c \frac{1}{x} - K}_{\Leftrightarrow y^* = b_1 x^* + b_0} = \frac{c - Kx}{x} \\ &\Leftrightarrow y(x) = \frac{x}{dx + c}, \end{aligned}$$

with  $d = -K = b_0$  and  $c = b_1$ , the constant of proportionality.

# Warning about linearizing transformations

A simple linear regression does **not directly** model non-linear relations between  $x$  and  $y$ . It may model a linear relation between **transformed** variables.

Transformations of the response variable  $y$  have a major impact on the fit: the scale of residuals is changed.

Concepts that depend on the scale of  $y$  values, such as *SQRE* and  $R^2$ , are **not directly comparable**, with or without a transformation of the response variable.

**Note:** Linearizing, obtaining the parameter values  $b_0$  and  $b_1$  for the regression line and then undoing the linearizing transformation does **not** give the same parameter values as would result from directly minimising the sum of squared residuals on the non-linear relation, using a **Non-linear Regression**.



# Multiple Linear Regression

It may be necessary to have **more than one predictor** to model the response variable of interest.

## Example 7: Antoci dataset

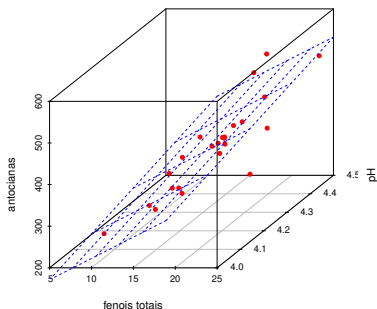
In a study of an experimental population of clones of the Tinta Francisca grape variety, carried out in Tabuaço in 2003, the following variables were observed on 24 grapevines:

- **anthocyan** (variable `antoci`, in  $mg/dm^3$ );
- **total phenols** (variable `fentot`);
- **pH** (variable `pH`).

We seek to study the relation between the anthocyan content (response variable) and the content of total phenols and pH.

## The scatterplot for Example 7

The  $n = 24$  observations of three variables produce a 24-point scatterplot in  $\mathbb{R}^3$ , which seems to be well approximated by a plane. The scatterplot was obtained using command `scatterplot3d`, from the R package with the same name.



The alternative `rggobi` package, permits the use of the `Ggobi` software and is a powerful tool for the visualization of 3-dimensional plots.

## Planes in $\mathbb{R}^3$

Any plane in  $\mathbb{R}^3$ , on the  $x_0y_0z$  system of axes, has an equation

$$Ax + By + Cz + D = 0 .$$

In our context, and associating:

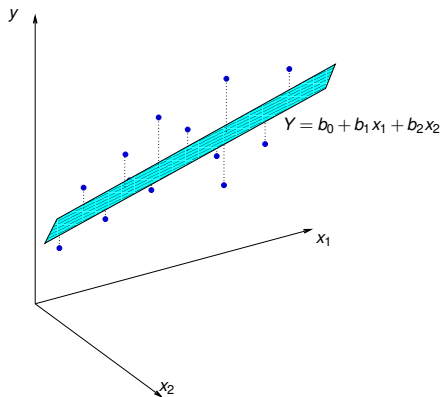
- the vertical axis ( $z$ ) with the response variable  $y$ ;
- axis  $x$  with one predictor,  $x_1$ ;
- the third axis ( $y$ ) with the other predictor,  $x_2$ ,

The equation becomes (if  $C \neq 0$ , i.e., for **non-vertical planes**):

$$\begin{aligned} Ax_1 + Bx_2 + Cy + D = 0 &\Leftrightarrow Cy = -D - Ax_1 - Bx_2 \\ &\Leftrightarrow y = -\frac{D}{C} - \frac{A}{C}x_1 - \frac{B}{C}x_2 \\ &\Leftrightarrow y = b_0 + b_1x_1 + b_2x_2 \end{aligned}$$

This equation **extends the straight line equation to 2 predictors**.

## Multiple linear regression ( $p=2$ predictors)



$y = b_0 + b_1 x_1 + b_2 x_2$  is the equation of a **plane** in  $\mathbb{R}^3$  ( $x_1 \geq 0, x_2 \geq 0, y \geq 0$ ).

The equation has **3 parameters**:  $b_0$ ,  $b_1$  and  $b_2$ . It can be **fitted with the same Least Squares criterion** used in a simple linear regression: **minimise SQRE**.

# The general case: $p$ predictors

We seek to model a **response variable**,  $y$ , based on  **$p$  predictors**,  $x_1, x_2, \dots, x_p$ . We have  $n$  observations on those  $p+1$  variables:

$$\left\{ (x_{1(i)}, x_{2(i)}, \dots, x_{p(i)}, y_i) \right\}_{i=1}^n .$$

**Problem:** The standard representation can no longer be visualised when  $p > 2$ , since the observations define an  **$n$ -point scatterplot in the space  $\mathbb{R}^{p+1}$** .

The main traits of the **standard representation** are:

- **$p+1$  axes** – one for each **variable**.
- **$n$  points** – one for each observed **individual** (experimental unit).
- An  **$n$ -point scatterplot in  $(p+1)$ -dimensional space**.

# Multiple linear regression: the fitted hyperplane

We assume that the observed values of  $y$  have an underlying trend given by a linear (affine) combination of the  $p$  predictor variables:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p .$$

This is the equation of a **hyperplane** in  $\mathbb{R}^{p+1}$ .

The **criterion** used to fit the hyperplane to the  $n$ -point scatterplot in  $\mathbb{R}^{p+1}$  is that of **minimising the Sum of Squared Residuals**, that is, choosing the  $p+1$  parameters  $\{b_j\}_{j=0}^p$  that minimise:

$$SQRE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where  $y_i$  are the observed values of the response variable and

$$\hat{y}_i = b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \dots + b_p x_{p(i)}$$

are the corresponding values fitted using the hyperplane equation.

## Two approaches to obtaining the fitted parameters

To obtain the parameters that define the best-fitting hyperplane it is possible to use two approaches:

- **analytic**; or
- **geometric**.

In both approaches, the use of a vector-matrix notation is crucial.

**No simple formulas exist**, as was the case in a simple linear regression, **for each individual parameter  $b_j$** . But it is possible to obtain a **single matrix formula to obtain all  $p+1$  model parameters at once**.

We shall follow the **geometric approach**.

# An alternative representation: the space of variables

The standard representation of the  $n$  observations of  $y$  and the  $p+1$  predictors, in  $\mathbb{R}^{p+1}$ , is not the only possible one.

There is an **alternative representation of the data, that merges geometric concepts and statistical concepts.**

The  $n$  observations of  $y$  define a vector with  $n$  coordinates, i.e., **in  $\mathbb{R}^n$** :

$$\vec{y} = (y_1, y_2, y_3, \dots, y_n)^t.$$

Likewise, the  $n$  observations of any given predictor variable define a vector **in  $\mathbb{R}^n$** :

$$\vec{x}_j = (x_{j(1)}, x_{j(2)}, x_{j(3)}, \dots, x_{j(n)})^t \quad (j = 1, 2, \dots, p).$$

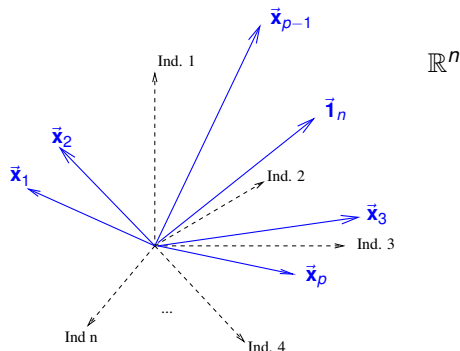
Thus, we can represent the variables as **points/vectors in  $\mathbb{R}^n$** .



# The representation in $\mathbb{R}^n$

In this **alternative representation**,

- each **axis** corresponds to an observed **individual**;
- each **vector** corresponds to a **variable**.



The **vector of  $n$  ones**, represented by  $\vec{i}_n$ , is also a vector in  $\mathbb{R}^n$ .

# The vector of fitted values

The  $n$  fitted values  $\hat{y}_i$  also define a vector in  $\mathbb{R}^n$ ,  $\vec{\hat{y}}$ :

$$\begin{aligned}\vec{\hat{y}} &= \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 x_{1(1)} + b_2 x_{2(1)} + \dots + b_p x_{p(1)} \\ b_0 + b_1 x_{1(2)} + b_2 x_{2(2)} + \dots + b_p x_{p(2)} \\ b_0 + b_1 x_{1(3)} + b_2 x_{2(3)} + \dots + b_p x_{p(3)} \\ \dots \\ b_0 + b_1 x_{1(n)} + b_2 x_{2(n)} + \dots + b_p x_{p(n)} \end{bmatrix} \\ &= b_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + b_1 \begin{bmatrix} x_{1(1)} \\ x_{1(2)} \\ x_{1(3)} \\ \vdots \\ x_{1(n)} \end{bmatrix} + \dots + b_p \begin{bmatrix} x_{p(1)} \\ x_{p(2)} \\ x_{p(3)} \\ \vdots \\ x_{p(n)} \end{bmatrix} \\ &= b_0 \vec{\mathbf{1}}_n + b_1 \vec{\mathbf{x}}_1 + b_2 \vec{\mathbf{x}}_2 + \dots + b_p \vec{\mathbf{x}}_p\end{aligned}$$

The vector  $\vec{\hat{y}}$  is a linear combination of the vectors  $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \dots, \vec{\mathbf{x}}_p$

# The model matrix $\mathbf{X}$

The vector  $\vec{\hat{y}}$  of fitted values can also be written as a product of a matrix  $\mathbf{X}$ , whose columns are the vectors  $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p$ .

## The model matrix $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1(1)} & x_{2(1)} & \cdots & x_{p(1)} \\ 1 & x_{1(2)} & x_{2(2)} & \cdots & x_{p(2)} \\ 1 & x_{1(3)} & x_{2(3)} & \cdots & x_{p(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1(n)} & x_{2(n)} & \cdots & x_{p(n)} \end{bmatrix}$$

$\underbrace{\hspace{1.5cm}}_{=\vec{\mathbf{1}}_n} \quad \underbrace{\hspace{1.5cm}}_{=\vec{\mathbf{x}}_1} \quad \underbrace{\hspace{1.5cm}}_{=\vec{\mathbf{x}}_2} \quad \cdots \quad \underbrace{\hspace{1.5cm}}_{=\vec{\mathbf{x}}_p}$

The model matrix  $\mathbf{X}$  has size  $n \times (p + 1)$ .

# The matrix products $\mathbf{X}\vec{a}$

Products of the form  $\mathbf{X}\vec{a}$  are **linear combinations of the columns of matrix  $\mathbf{X}$** :

$$\begin{aligned}\mathbf{X}\vec{a} &= \begin{bmatrix} 1 & x_{1(1)} & x_{2(1)} & \cdots & x_{p(1)} \\ 1 & x_{1(2)} & x_{2(2)} & \cdots & x_{p(2)} \\ 1 & x_{1(3)} & x_{2(3)} & \cdots & x_{p(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1(n)} & x_{2(n)} & \cdots & x_{p(n)} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \\ &= \begin{bmatrix} a_0 + a_1 x_{1(1)} + a_2 x_{2(1)} + \dots + a_p x_{p(1)} \\ a_0 + a_1 x_{1(2)} + a_2 x_{2(2)} + \dots + a_p x_{p(2)} \\ a_0 + a_1 x_{1(3)} + a_2 x_{2(3)} + \dots + a_p x_{p(3)} \\ \dots \\ a_0 + a_1 x_{1(n)} + a_2 x_{2(n)} + \dots + a_p x_{p(n)} \end{bmatrix} \\ &= a_0 \vec{1}_n + a_1 \vec{x}_1 + a_2 \vec{x}_2 + \dots + a_p \vec{x}_p\end{aligned}$$

The **vector  $\vec{y}$**  can be written in this way:  $\vec{y} = \mathbf{X}\vec{b}$ , for some vector of (as of yet unknown) coefficients  $\vec{b} \in \mathbb{R}^{p+1}$ .

# The model matrix $\mathbf{X}$ and its column-space

- The set of **all linear combinations** of a set of vectors is called the **subspace spanned** by those vectors.
- The subspace spanned by the columns of the model matrix  $\mathbf{X}$  is called the **column-space of matrix  $\mathbf{X}$** ,  $\mathcal{C}(\mathbf{X})$ .
- The vector  $\vec{\hat{\mathbf{y}}}$  belongs to the subspace  $\mathcal{C}(\mathbf{X})$  (the vectors  $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p$  are columns of  $\mathbf{X}$  and  $\vec{\hat{\mathbf{y}}} = b_0 \vec{\mathbf{1}}_n + b_1 \vec{\mathbf{x}}_1 + b_2 \vec{\mathbf{x}}_2 + \dots + b_p \vec{\mathbf{x}}_p$ ).
- $\mathcal{C}(\mathbf{X})$  is a subspace of  $\mathbb{R}^n$  ( $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$ ), but of **dimension  $p+1$**  (assuming the columns of  $\mathbf{X}$  are **linearly independent**, that is, if none of those vectors can be written as a linear combination of the others).
- Any linear combination of the columns of matrix  $\mathbf{X}$ , that is, **any element of  $\mathcal{C}(\mathbf{X})$**  can be written as  $\mathbf{X}\vec{\mathbf{a}}$ , where  $\vec{\mathbf{a}} = (a_0, a_1, a_2, \dots, a_p)$  is the vector of coefficients of the linear combination.

# The parameters

- Each possible choice of coefficients  $\vec{\mathbf{a}} = (a_0, a_1, a_2, \dots, a_p)$  corresponds to a point/vector of subspace  $\mathcal{C}(\mathbf{X})$ .
- That choice of coefficients is **unique** if the columns of  $\mathbf{X}$  are **linearly independent**, that is, if there is no linear dependence (**multi-collinearity**) between the variables  $\vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p, \vec{\mathbf{1}}_n$ .
- One of the points/vectors in the subspace is the linear combination given by the vector of coefficients  $\vec{\mathbf{b}} = (b_0, b_1, \dots, b_p)$ , that minimises:

$$SQRE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where the  $y_i$  are the observed values of the response variable and  $\hat{y}_i = b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \dots + b_p x_{p(i)}$  are the **fitted values**. This is the linear combination that we seek.

How do we identify this point/vector?

# Geometry

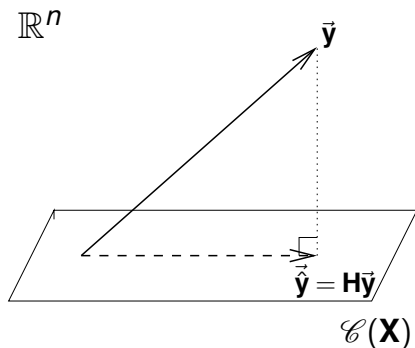
Let us use geometric arguments.

- We have a vector of  $n$  observations of  $\vec{y}$  which belongs to  $\mathbb{R}^n$  but, in general, does not belong to the subspace  $\mathcal{C}(\mathbf{X})$ .
- We wish to approximate this vector by another vector,  $\vec{\hat{y}} = b_0 \vec{1}_n + b_1 \vec{x}_1 + \dots + b_p \vec{x}_p$ , which belongs to the subspace  $\mathcal{C}(\mathbf{X})$ .
- Let us approximate the vector of observations  $\vec{y}$  by the vector  $\vec{\hat{y}}$  in subspace  $\mathcal{C}(\mathbf{X})$  that is closest to  $\vec{y}$ .

## SOLUTION:

Take the orthogonal projection of  $\vec{y}$  onto  $\mathcal{C}(\mathbf{X})$  :  $\vec{\hat{y}} = \mathbf{H}\vec{y}$ .

## The orthogonal projection of $\vec{y}$ onto $\mathcal{C}(\mathbf{X})$



The vector of  $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$  closest to vector  $\vec{y} \in \mathbb{R}^n$  is the vector  $\vec{\hat{y}}$  that results from orthogonally projecting  $\vec{y}$  onto  $\mathcal{C}(\mathbf{X})$ .

This orthogonal projection creates a **right triangle** in  $\mathbb{R}^n$ .



# The criterion minimises *SQRE*

Recall definitions regarding vectors:

- The **norm** (size) of vector  $\vec{\mathbf{x}} = (x_1, x_2, \dots, x_n)^t$  is  $\|\vec{\mathbf{x}}\| = \sqrt{\vec{\mathbf{x}}^t \vec{\mathbf{x}}} = \sqrt{\sum_{i=1}^n x_i^2}$ .
- The **distance** between two vectors  $\vec{\mathbf{x}}$  is  $\vec{\mathbf{y}}$  the norm of their difference:  
$$\text{dist}(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \|\vec{\mathbf{x}} - \vec{\mathbf{y}}\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}.$$

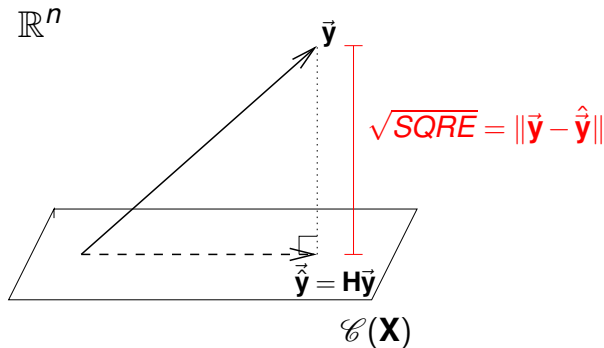
Choosing the vector  $\vec{\hat{\mathbf{y}}} \in \mathcal{C}(\mathbf{X})$  that minimises the distance to the vector of observations  $\vec{\mathbf{y}}$  means minimising the squared distance:

$$\text{dist}^2(\vec{\mathbf{y}}, \vec{\hat{\mathbf{y}}}) = \|\vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}\|^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \text{SQRE}.$$

In other words, **minimising the sum of squared residuals**.

The geometric concept is equivalent to the statistical criterion used to fit the parameters in a Linear Regression.

## SQRE in the orthogonal projection



The squared distance between  $\vec{y}$  and  $\hat{\vec{y}}$  is  $SQRE$ , the sum of squared residuals.

# Orthogonal projections

The orthogonal projection of a vector  $\vec{y} \in \mathbb{R}^n$  onto the subspace  $\mathcal{C}(\mathbf{X})$  spanned by the (linearly independent) columns of  $\mathbf{X}$  results from pre-multiplying  $\vec{y}$  by the **matrix of orthogonal projections onto  $\mathcal{C}(\mathbf{X})$** :

**Matrix of orthogonal projections onto  $\mathcal{C}(\mathbf{X})$**

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t.$$

The **matrices of orthogonal projections  $\mathbf{P}$**  onto some subspace of  $\mathbb{R}^n$  are  $n \times n$  **matrices** that are:

- **symmetric** (that is,  $\mathbf{P}^t = \mathbf{P}$ ); and
- **idempotent** (that is,  $\mathbf{P}\mathbf{P} = \mathbf{P}$ ).

Matrix  $\mathbf{H}$  has these properties (Exercise RL 11: confirm!).

# Orthogonal projections in the context of a MLR

In the context of a multiple linear regression, we have:

$$\begin{aligned} \vec{\hat{y}} &= \mathbf{H}\vec{y} \\ \Leftrightarrow \vec{\hat{y}} &= \underbrace{\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{y}}_{=\vec{b}} \end{aligned}$$

The linear combination of the vectors  $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p$  that generates the vector closest to  $\vec{y}$  has coefficients given by the elements of vector  $\vec{b}$ :

## The vector of fitted parameters

$$\vec{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{y} .$$

# The three Sums of Squares

Recall the three Sums of Squares:

**SQRE** The Residual Sum of Squares:

$$SQRE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 .$$

**SQT** The Total Sum of Squares:

$$SQT = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 .$$

**SQR** The Regression Sum of Squares:

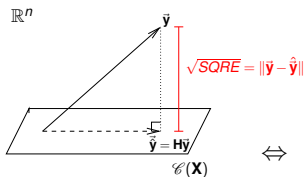
$$SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2 .$$

# Pythagoras and Linear Regression

**Pythagoras' Theorem** is valid in any Euclidean space  $\mathbb{R}^n$ .

The right triangle on slide 82 produces the following relation:

$$\|\vec{y}\|^2 = \|\hat{y}\|^2 + \|\vec{y} - \hat{y}\|^2$$



$$\Leftrightarrow \sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{= SQRE}$$

$$\Leftrightarrow \sum_{i=1}^n y_i^2 - n\bar{y}^2 = \sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2 + SQRE$$

$$\Leftrightarrow SQT = SQR + SQRE$$

# Revisiting Pythagoras

The fundamental relation of Linear Regressions ( $SQT = SQR + SQRE$ ) results from applying Pythagoras' Theorem. But it was necessary to subtract  $n\bar{y}^2$  from both sides of the equation. A different right triangle is statistically more interesting.

Define the **centred vector**,  $\vec{y}^c$ , the generic element of which is the deviation of each  $y_i$  from the mean:  $y_i - \bar{y}$ .

$$\vec{y}^c = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix} = \vec{y} - (\bar{y})\vec{\mathbf{1}}_n.$$

The norm of this vector is  $\|\vec{y}^c\| = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} = \sqrt{SQT}$ .

## Revisiting Pythagoras (cont.)

The orthogonal projection of vector  $\vec{y}^c$  onto the subspace  $\mathcal{C}(\mathbf{X})$  produces the vector:

$$\begin{aligned}\mathbf{H}\vec{y}^c &= \mathbf{H}[\vec{y} - (\bar{y}) \cdot \vec{\mathbf{1}}_n] \\ \Leftrightarrow \mathbf{H}\vec{y}^c &= \mathbf{H}\vec{y} - (\bar{y}) \cdot \mathbf{H}\vec{\mathbf{1}}_n \\ \Leftrightarrow \mathbf{H}\vec{y}^c &= \vec{\hat{y}} - (\bar{y}) \cdot \vec{\mathbf{1}}_n\end{aligned}$$

since  $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$ , because vector  $\vec{\mathbf{1}}_n$  already belongs to the subspace  $\mathcal{C}(\mathbf{X})$ , and so remains invariant when projected onto that same subspace – see Exercise 11.

The vector  $\mathbf{H}\vec{y}^c$  has the generic element  $\hat{y}_i - \bar{y}$ . Its norm is:

$$\|\mathbf{H}\vec{y}^c\| = \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2} = \sqrt{SQR}.$$



## Revisiting Pythagoras (cont.)

The distance between vector  $\vec{\mathbf{y}}^c$  and its orthogonal projection onto  $\mathcal{C}(\mathbf{X})$  continues to be  $\sqrt{SQRE}$ :

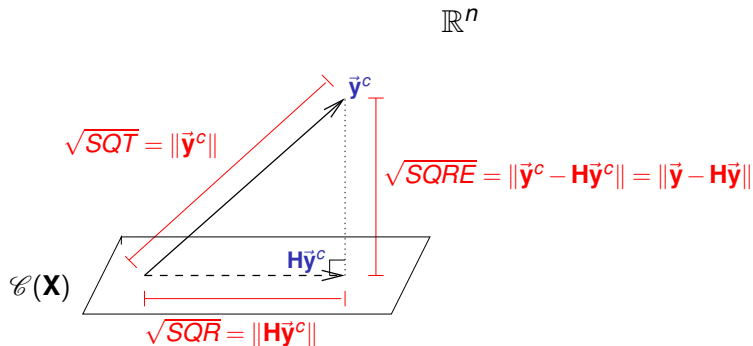
$$\begin{aligned}\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c &= [\vec{\mathbf{y}} - \cancel{\bar{y}\vec{\mathbf{1}}_n}] - [\vec{\hat{\mathbf{y}}} - \cancel{\bar{y}\vec{\mathbf{1}}_n}] \\ \Leftrightarrow \vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c &= \vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}\end{aligned}$$

and so:

$$\|\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c\| = \|\vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}\| = \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2} = \sqrt{SQRE}.$$

## Revisiting Pythagoras (cont.)

The fundamental formula of Linear Regression,  $SQT = SQR + SQRE$ , results from a direct application of the Pythagorean Theorem to the triangle defined by  $\vec{y}^c$  and its orthogonal projection onto  $\mathcal{C}(\mathbf{X})$ .



# Pythagoras and the Coefficient of Determination

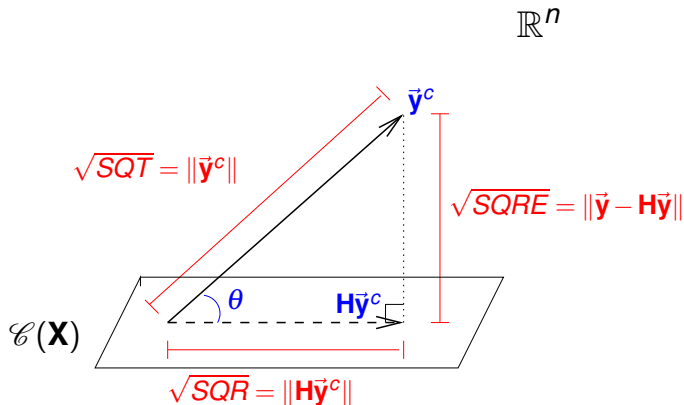
Another important connection between the geometry of space  $\mathbb{R}^n$  and Linear Regression exists:

The **coefficient of determination**  $R^2 = \frac{SQR}{SQT}$  is the squared cosine of the angle between the centred vector of observations of the response variable,  $\vec{\mathbf{y}}^c$ , and its orthogonal projection onto subspace  $\mathcal{C}(\mathbf{X})$ :

$$\cos^2(\theta) = \frac{SQR}{SQT} = R^2 ,$$

where  $\theta$  is the angle between vectors  $\vec{\mathbf{y}}^c$  and  $\mathbf{H}\vec{\mathbf{y}}^c$ .

# Pythagoras and Coefficients of Determination (cont.)



The Coefficient of Determination in a Linear Regression,  $R^2 = \frac{SQR}{SQT}$ , is the squared cosine of the angle between  $\vec{y}^c$  and  $H\vec{y}^c$ .

# Properties of the Coefficient of Determination

The geometric approach confirms that, in a Multiple Linear Regression too, the Coefficient of Determination has well-known properties:

- $R^2$  can take values between 0 and 1.
- The closer  $R^2$  is to 1, the smaller the angle  $\theta$ , and so the better the match between the (centred) vector of observations  $\bar{\mathbf{y}}^c$  and its fit onto  $\mathcal{C}(\mathbf{X})$ .
- If  $R^2 \approx 0$ , vector  $\bar{\mathbf{y}}^c$  is almost perpendicular to the subspace  $\mathcal{C}(\mathbf{X})$  where it is being approximated and the projection will almost nullify all the elements of the projected vector, that is,  $\hat{y}_i - \bar{y} \approx 0$ . The result is of poor quality: we lose almost all the variability in the values of  $\hat{y}_i \approx \bar{y}$ .

# Properties of models with an intercept

$\mathcal{C}(\mathbf{X})$  contains the vector  $\vec{\mathbf{1}}_n$  with  $n$  ones. Hence,  $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$ , because the projection of any vector onto a subspace that already contains it leaves the vector invariant. Therefore, (see also Exercise 11):

- The mean of the observed and fitted values of  $y$  is the same:

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \vec{\mathbf{1}}_n^t \vec{\hat{y}} = \frac{1}{n} \vec{\mathbf{1}}_n^t \mathbf{H}\vec{y} = \frac{1}{n} \vec{\mathbf{1}}_n^t \mathbf{H}^t \vec{y} = \frac{1}{n} (\mathbf{H}\vec{\mathbf{1}}_n)^t \vec{y} = \frac{1}{n} \vec{\mathbf{1}}_n^t \vec{y} = \bar{y}$$

- The sum of residuals is zero:

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = \bar{y} - \bar{\hat{y}} = 0.$$

- In  $\mathbb{R}^{p+1}$ , the fitted hyperplane contains the centre of gravity of the observed  $n$ -point scatterplot:  $\bar{y} = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \dots + b_p \bar{x}_p$ .

We have already seen that  $\bar{y} = \bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i$ . But  $\frac{1}{n} \sum_{i=1}^n \hat{y}_i =$

$$\frac{1}{n} \sum_{i=1}^n (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \dots + b_p x_{p(i)}) = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \dots + b_p \bar{x}_p$$

## The coefficients $b_j$

The vector of parameters fitted by the least squares method,  $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}}$ , generates  $n$  fitted values:

$$\begin{aligned}\vec{\hat{\mathbf{y}}} &= \mathbf{H}\vec{\mathbf{y}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}} = \mathbf{X}\vec{\mathbf{b}} \\ \Leftrightarrow \hat{y}_i &= b_0 + b_1x_{1(i)} + \dots + b_px_{p(i)}, \quad \forall i.\end{aligned}$$

The units of measurement:

- of  $b_0$  are the same as those of  $y$  (and of  $\hat{y}$ ).
- of the parameters  $b_j$  that multiply variables ( $j \neq 0$ ) are the ratio of the units of  $y$  over the units of the corresponding  $x_j$ .

The coefficients  $\{b_j\}_{j=1}^p$  of the predictors can be interpreted as the mean difference in  $y$ , associated with increasing predictor  $x_j$  by one unit, while keeping constant all remaining predictors.

# Residuals

The **units of measurement** of the **residuals**  $e_i = y_i - \hat{y}_i$  are the same as those of  $y$ :

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1(i)} + \dots + b_p x_{p(i)}) \quad , \quad \forall i$$
$$\Leftrightarrow \vec{e} = \vec{y} - \vec{\hat{y}} = \vec{y} - \mathbf{H}\vec{y} \quad ,$$

The vector of **residuals**,  $\vec{e}$ , can also be obtained by **pre-multiplying** vector  $\vec{y}$  by the matrix  $\mathbf{I} - \mathbf{H}$ , where  $\mathbf{I}$  is the  $n \times n$  identity matrix:


$$\vec{e} = \vec{y} - \mathbf{H}\vec{y} = (\mathbf{I} - \mathbf{H})\vec{y} \quad ,$$

Matrix  $\mathbf{I} - \mathbf{H}$  is **symmetric** and **idempotent**, hence it too is a matrix of orthogonal projections. It projects onto the subspace of  $\mathbb{R}^n$  of the vectors that are orthogonal to all vectors of  $\mathcal{C}(\mathbf{X})$ , a subspace called the **orthogonal complement of  $\mathcal{C}(\mathbf{X})$**  and denoted  $\mathcal{C}(\mathbf{X})^\perp$ .

Vector  $\vec{e}$  is the **orthogonal projection** of  $\vec{y}$  onto  $\mathcal{C}(\mathbf{X})^\perp$ .



## Multiple Linear Regressions in

The command `lm` also fits Multiple Linear Regressions in . The response variable  $y$  and the predictors  $x_1, \dots, x_p$  are defined using a formula similar to that used in simple linear regressions.

E.g., if  $y$  is the response variable and  $x_1$ ,  $x_2$  and  $x_3$  are three predictors, the formula that specifies the relation will be:

$$y \sim x_1 + x_2 + x_3$$

### R command fitting a multiple linear regression

```
> lm ( y ~ x1 + x2 + x3 + ... + xp, data=dados)
```

The command returns the vector  $\vec{b}$  with the fitted values of the  $p + 1$  model parameters,  $b_0, b_1, \dots, b_p$ .

# An example of MLR in R

We illustrate a Multiple Linear Regression in R with a famous dataset: the [iris data of Anderson/Fisher](#), available in the *data frame* `iris`.

```
> help(iris)
```

```
iris                                package:datasets                                R Documentation
```

```
Edgar Anderson's Iris Data
```

```
Description:
```

```
This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.
```

## MLR example (cont.)



Figura: iris setosa



Figura: iris versicolor



Figura: iris virginica

## MLR example (cont.)

An initial inspection of the data can be carried out with command `head`, that shows the first rows of the *data frame*:

```
> head(iris)
```

```
 Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1           3.5           1.4           0.2  setosa
2           4.9           3.0           1.4           0.2  setosa
3           4.7           3.2           1.3           0.2  setosa
4           4.6           3.1           1.5           0.2  setosa
5           5.0           3.6           1.4           0.2  setosa
6           5.4           3.9           1.7           0.4  setosa
```

The main indicators are given by the command `summary`:

```
> summary(iris)
```

```
 Sepal.Length      Sepal.Width      Petal.Length      Petal.Width      Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa   :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Note that the fifth column is a **factor**. It will, for now, be ignored.

# A descriptive Multiple Linear Regression in (cont.)

A linear regression model to predict the response variable petal width was fitted, using the petal length and both sepal measurements (width and length) as predictors, **ignoring species**.

## Multiple LR - iris data

```
> iris2.lm <- lm(Petal.Width ~ Petal.Length + Sepal.Length +  
+               Sepal.Width , data=iris)  
> iris2.lm  
(...)  
Coefficients:  
 (Intercept)  Petal.Length  Sepal.Length  Sepal.Width  
    -0.2403         0.5241        -0.2073         0.2228
```

The fitted hyperplane in  $\mathbb{R}^4$  ( $\mathbb{R}^{p+1}$ ) is:

$$PW = -0.2403 + 0.5241 PL - 0.2073 SL + 0.2228 SW$$

## Confirming the formula (cont.)

Let us confirm the formula for the parameters fitted by the least squares method. The command `model.matrix` returns matrix  $\mathbf{X}$ .

```
> X <- model.matrix(iris2.lm)
> X
```

	(Intercept)	Petal.Length	Sepal.Length	Sepal.Width
1	1	1.4	5.1	3.5
2	1	1.4	4.9	3.0
3	1	1.3	4.7	3.2
4	1	1.5	4.6	3.1
5	1	1.4	5.0	3.6
6	1	1.7	5.4	3.9
7	1	1.4	4.6	3.4
8	1	1.5	5.0	3.4
[...]				
149	1	5.4	6.2	3.4
150	1	5.1	5.9	3.0

## Confirming the formula (cont.)

The necessary R commands for the matrix operations in  $\vec{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{y}$  are:

- `t(A)` indicates the **transpose of matrix A**
- `A %*% B` indicates the **matrix product** of A and B.
- `solve(A)` computes the **matrix inverse** of A.

```
> y <- iris$Petal.Width  
> b <- solve( t(X) %*% X ) %*% ( t(X) %*% y )  
> b
```

```
          [,1]  
(Intercept) -0.2403074  
Petal.Length  0.5240831  
Sepal.Length -0.2072661  
Sepal.Width   0.2228285
```

The values on slide 101 are confirmed.

# Models and submodels

## Submodels

Given a multiple linear regression model, with equation

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p ,$$

we call a linear regression with only **some** of these predictors a **submodel**.

For example, a **simple** linear regression

$$Petal.Width = b_0 + b_1Petal.Length$$

is a **submodel** of the multiple linear regression that was fitted above,

$$Petal.Width = b_0 + b_1Petal.Length + b_2Sepal.Length + b_3Sepal.Width$$

**Warning:** A submodel (S) cannot have predictors that were not used in the complete (C), or full, model. The response variable must be the same.



## $R^2$ in submodels

Coefficients of Determination in submodels:  $R_S^2 \leq R_C^2$

The  $R_S^2$  of a submodel cannot be larger than the  $R_C^2$  of the full model.

The column-space of a submodel is contained in the column-space of the full model:  $\mathcal{C}(\mathbf{X}_S) \subseteq \mathcal{C}(\mathbf{X}_C)$ . Hence, the angle between  $\vec{y}$  and  $\vec{y}_S \in \mathcal{C}(\mathbf{X}_S)$  cannot be smaller than the angle between  $\vec{y}$  and  $\vec{y}_C \in \mathcal{C}(\mathbf{X}_C)$ , since  $\vec{y}_S$  also belongs to  $\mathcal{C}(\mathbf{X}_C)$ .

For the model in slide 101:  $R^2 = 0.9379$ .

For the simple linear regression with predictor `Petal.Length`:  $R^2 = 0.9271$ .

### Still the iris example

```
> summary(iris2.lm)$r.sq
[1] 0.9378503
> iris.lm <- lm(Petal.Width ~ Petal.Length, data = iris)
> summary(iris.lm)$r.sq
[1] 0.9271098
```

# Equation of submodels

The fitted parameter values are not the same

The fitted equation in a submodel **is not** the corresponding part of the fitted equation for the full model.

## Again the iris example

```
> coef(iris.lm)
```

```
(Intercept) Petal.Length  
-0.3630755    0.4157554
```

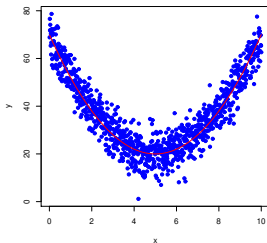
```
> coef(iris2.lm)
```

```
(Intercept) Petal.Length Sepal.Length Sepal.Width  
-0.2403074    0.5240831   -0.2072661    0.2228285
```

# Polynomial regression

A specific case of non-linear relation, even if only with a single predictor, may be easily studied using multiple linear regression: the case of polynomial relations between  $y$  and one or more predictors.

Imagine an underlying parabolic relation between a response variable  $y$  and a single predictor  $x$  given by:

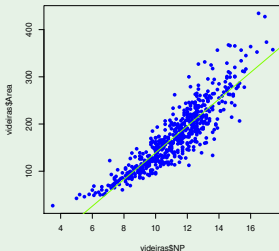


# Polynomial Regression - Example

## Example 5 – Vine leaves

Consider the dataset with measurements on  $n=600$  vine leaves.

This is the scatterplot for **areas** vs. **length of the main vein**, with the regression line on top.



There is an underlying curvature. Maybe a 2d degree polynomial?

## Polynomial regression- Example (cont.)

A parabola, with equation

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2,$$

can be fitted as a linear regression of  $y$  on 2 predictors  $X_1 = X$  and  $X_2 = X^2$ :

```
> videiras.lm2 <- lm( Area ~ NP + I(NP^2) , data=videiras )  
> videiras.lm2
```

Call:

```
lm(formula = Area ~ NP + I(NP^2), data = videiras)
```

Coefficients:

(Intercept)	NP	I(NP^2)
7.5961	-0.2172	1.2941

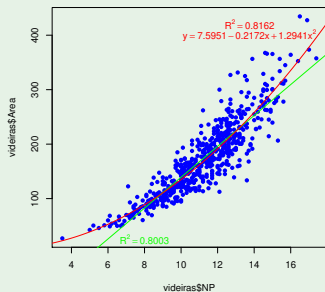
```
> summary( videiras.lm2 )$r.sq
```

```
[1] 0.8161632
```

The equation of the fitted parabola is  $y = 7.5961 - 0.2172x + 1.2941x^2$ . The value  $R^2 = 0.8162$  indicates that **some 82% of the observed variability in leaf surface areas is accounted for by the quadratic regression** (here,  $y$  was not transformed).

# Polynomial regression - Example (cont.)

## The fitted parabola



The equation of the fitted line is  $y = -144.15 + 28.34x$ , confirming that the fitted equation of a submodel (in this case, the regression line) **is not** the relevant part of the equation fitted for the full model (in this case, the parabolic model).

## Polynomial regressions (cont.)

A similar reasoning applies with polynomials of any degree, and for any number of predictors. Two examples:

- A  $p$ -th degree polynomial on a single variable:

$$Y = \beta_0 + \beta_1 \underbrace{x}_{=x_1} + \beta_2 \underbrace{x^2}_{=x_2} + \beta_3 \underbrace{x^3}_{=x_3} + \dots + \beta_p \underbrace{x^p}_{=x_p}$$

- A second degree polynomial with two variables:

$$Y = \beta_0 + \beta_1 \underbrace{x}_{=x_1} + \beta_2 \underbrace{x^2}_{=x_2} + \beta_3 \underbrace{z}_{=x_3} + \beta_4 \underbrace{z^2}_{=x_4} + \beta_5 \underbrace{xz}_{=x_5}$$

## Linear regression - Inference

- So far, linear regression was only used as a **descriptive method**. If the  $n$  observations were the totality of the population of interest, there would be little to add.
- But the  $n$  observations are often just a **random sample** from a larger population.
- A fitted hyperplane,  $y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$ , based on any one sample is merely an **estimate** of a **population hyperplane**

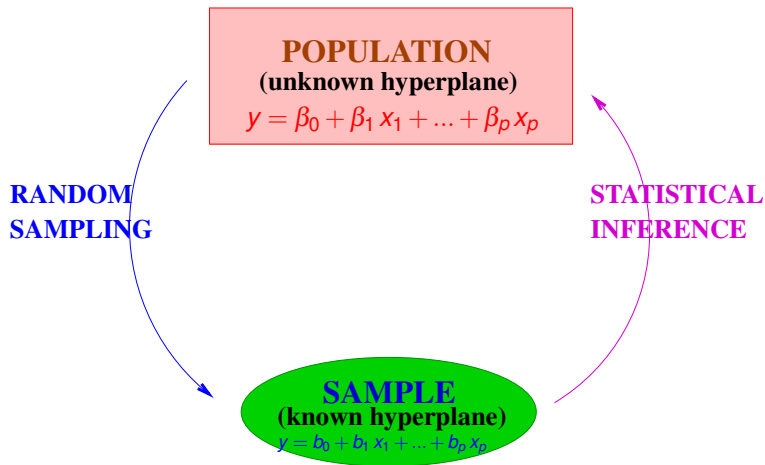
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p .$$

Other samples would give rise to different fitted hyperplanes.

- The issue of **statistical inference** becomes relevant.



# Statistical inference in Linear Regression



# MODEL - Linear Regression

In order to make inference about the population hyperplane possible, we must make **additional assumptions**.

$Y$  – **random** response variable.

$x_1, \dots, x_p$  – **non-random** predictor variables (controlled by the experimenter or the model will be **conditional** on the observed values of  $x_1, \dots, x_p$ )

The model will be fitted based on:

$\{(x_{1(i)}, x_{2(i)}, \dots, x_{p(i)}, Y_i)\}_{i=1}^n$  –  $n$  sets of **independent** observations of the variables  $x_1, x_2, \dots, x_p$  and  $Y$ , on  $n$  **experimental units**.

# LR MODEL – Linearity

We also assume that there is an **underlying relation between  $Y$  and  $x_1, x_2, \dots, x_p$ , that is linear (affine)**, with random variability around that trend relation. This variability is represented by a **random error  $\varepsilon$** . For all  $i = 1, \dots, n$ :

$$\begin{array}{ccccccccccc} Y_i & = & \beta_0 & + & \beta_1 & x_{1(i)} & + & \dots & + & \beta_p & x_{p(i)} & + & \varepsilon_i \\ \downarrow & & \downarrow & & \downarrow & \downarrow & & & & \downarrow & \downarrow & & \downarrow \\ \text{r.v.} & & \text{ct.} & & \text{ct.} & \text{ct.} & & & & \text{ct.} & \text{ct.} & & \text{r.v.} \end{array}$$

# Linear Regression MODEL – Random errors

We also **assume that the random errors  $\varepsilon_j$** :

- Have an **expected value** (mean value) of **zero**:

$$E[\varepsilon_j] = 0, \quad \forall i = 1, \dots, n$$

(this is not a restrictive assumption).

- **Have a Normal distribution** (this is restrictive, but fairly general).
- **Variance homogeneity**: all errors have the same variance

$$V[\varepsilon_j] = \sigma^2, \quad \forall i = 1, \dots, n$$

(restrictive, but convenient).

- Are **independent random variables** (r.v.)  
(restrictive, but convenient).

# The Linear Model

The model for inferential purposes in a linear regression is therefore:

## The Linear Model

- 1  $Y_i = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \dots + \beta_p x_{p(i)} + \varepsilon_i, \quad \forall i = 1, \dots, n.$
- 2  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \forall i = 1, \dots, n.$
- 3  $\{\varepsilon_i\}_{i=1}^n$  are independent r.v..

**NOTE:** The random errors are independent and identically distributed (i.i.d.) random variables.

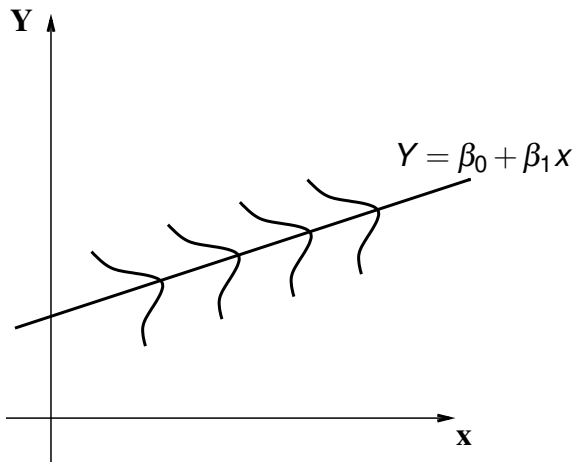
Given the model, the expected (mean) value of  $Y_i$ , conditional on the values  $x_1, x_2, \dots, x_p$  of the predictors, is:

$$\mu_i = E[Y_i | x_1, x_2, \dots, x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p.$$

**NOTE:**  $\beta_j$  ( $j \neq 0$ ) is the mean change in  $Y$ , associated with an increase of one unit in  $x_j$ , whilst keeping the remaining predictors constant.

# Simple Linear Regression MODEL

Illustrating the model in the case of a **simple** linear regression:



# Studying the model

A first inferential goal: the  $p+1$  model parameters,  $\beta_j$  ( $j = 0, 1, \dots, p$ ).

The fitted parameters  $\vec{b} = (b_0, b_1, b_2, \dots, b_p)$ , obtained applying the formula on slide 84 for a given sample, are estimates of those parameters.

In order to obtain confidence intervals and/or carry out hypothesis tests on the values of the population parameters  $\beta_j$ , we must:

- Define estimators  $\hat{\beta}_j$  for the population parameters;
- Determine their probability distributions (given the model);

The validity of the inference depends on the validity of the model assumptions.

# The matrix/vector notation

Studying the model (namely when there is more than one predictor) requires **appropriate tools** to deal with **random vectors**.

The model equations for the  $n$  observations (slide 117) may be written as a **single equation**, using vector/matrix notation:

$$\begin{array}{rccccccc} Y_1 & = & \beta_0 + \beta_1 x_{1(1)} + \beta_2 x_{2(1)} + \cdots + \beta_p x_{p(1)} & + & \varepsilon_1 \\ Y_2 & = & \beta_0 + \beta_1 x_{1(2)} + \beta_2 x_{2(2)} + \cdots + \beta_p x_{p(2)} & + & \varepsilon_2 \\ Y_3 & = & \beta_0 + \beta_1 x_{1(3)} + \beta_2 x_{2(3)} + \cdots + \beta_p x_{p(3)} & + & \varepsilon_3 \\ \vdots & & \vdots & & \vdots \\ Y_n & = & \beta_0 + \beta_1 x_{1(n)} + \beta_2 x_{2(n)} + \cdots + \beta_p x_{p(n)} & + & \varepsilon_n \\ \underbrace{= \vec{Y}} & & \underbrace{= \mathbf{X}\vec{\beta}} & & \underbrace{= \vec{\varepsilon}} \end{array}$$



## Matrix/vector notation (cont.)

The  $n$  equations correspond to **a single vector equation**:

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon},$$

where:

$$\vec{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1(1)} & x_{2(1)} & \cdots & x_{p(1)} \\ 1 & x_{1(2)} & x_{2(2)} & \cdots & x_{p(2)} \\ 1 & x_{1(3)} & x_{2(3)} & \cdots & x_{p(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1(n)} & x_{2(n)} & \cdots & x_{p(n)} \end{bmatrix}, \quad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- $\vec{Y}$  and  $\vec{\epsilon}$  are **random** vectors,
- $\mathbf{X}$  is a **non-random** matrix and  $\vec{\beta}$  a **non-random** vector.

## The vector of estimators $\vec{\hat{\beta}}$

The **vector of estimators**  $\vec{\hat{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^t$  is defined from the equation for the vector of estimates  $\vec{\mathbf{b}}$  (slide 84), but replacing the vector  $\vec{\mathbf{y}}$  of observed values of  $y$  with the **random vector**  $\vec{\mathbf{Y}}$ .

### Least Squares parameter estimators

$$\vec{\hat{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{\mathbf{Y}} .$$

The resulting estimators are the **least squares estimators**.

Given the Linear Model, they are also maximum likelihood estimators.

## Tools for random vectors

Three **random vectors** have already been introduced:

- $\vec{Y}$  (the  $n$  observations of the response variable);
- $\vec{\epsilon}$  (the  $n$  random errors); and
- $\vec{\hat{\beta}}$  (the  $p+1$  estimators  $\hat{\beta}_j$ ).

We need **tools** to work with random vectors.

For any **random vector**  $\vec{Z} = (Z_1, Z_2, \dots, Z_k)^t$ , we define:

- The **expected vector** of  $\vec{Z}$ , with the **expected values** of each component:

$$\vec{\mu}_Z = E[\vec{Z}] = \begin{bmatrix} E[Z_1] \\ E[Z_2] \\ \vdots \\ E[Z_k] \end{bmatrix}.$$

If  $\mathbf{W}$  is a **random matrix**, we can also define  $E[\mathbf{W}]$  as the matrix of the expected values of each element.

## Tools for random vectors (cont.)

- the **variance-covariance matrix** of  $\vec{Z}$  has as elements the covariances for each pair of components:

$$V[\vec{Z}] = \begin{bmatrix} V[Z_1] & C[Z_1, Z_2] & C[Z_1, Z_3] & \dots & C[Z_1, Z_k] \\ C[Z_2, Z_1] & V[Z_2] & C[Z_2, Z_3] & \dots & C[Z_2, Z_k] \\ C[Z_3, Z_1] & C[Z_3, Z_2] & V[Z_3] & \dots & C[Z_3, Z_k] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C[Z_k, Z_1] & C[Z_k, Z_2] & C[Z_k, Z_3] & \dots & V[Z_k] \end{bmatrix}$$

This is necessarily a **symmetric matrix**.

## Properties of the expected vector

As in the case with random variables, so too the expected vector of a random vector  $\vec{\mathbf{Z}}_{k \times 1}$  has simple **properties**:

- If  $b$  is a **non-random** scalar,  $E[b\vec{\mathbf{Z}}] = b E[\vec{\mathbf{Z}}]$ .
- If  $\vec{\mathbf{a}}_{k \times 1}$  is a **non-random** vector,  $E[\vec{\mathbf{Z}} + \vec{\mathbf{a}}] = E[\vec{\mathbf{Z}}] + \vec{\mathbf{a}}$ .
- If  $\vec{\mathbf{a}}_{k \times 1}$  is a **non-random** vector,  $E[\vec{\mathbf{a}}^t \vec{\mathbf{Z}}] = \vec{\mathbf{a}}^t E[\vec{\mathbf{Z}}]$ .
- If  $\mathbf{B}_{m \times k}$  is a **non-random** matrix,  $E[\mathbf{B}\vec{\mathbf{Z}}] = \mathbf{B} E[\vec{\mathbf{Z}}]$ .

Also, the expected vector of the **sum of two random vectors** has the simple property:

- If  $\vec{\mathbf{Z}}_{k \times 1}$ ,  $\vec{\mathbf{U}}_{k \times 1}$  are **random** vectors,  $E[\vec{\mathbf{Z}} + \vec{\mathbf{U}}] = E[\vec{\mathbf{Z}}] + E[\vec{\mathbf{U}}]$ .

## Properties of the (co)variance matrix

- If  $b$  is a non-random scalar,  $V[b\vec{Z}] = b^2 V[\vec{Z}]$ .
- If  $\vec{a}_{k \times 1}$  is a non-random vector,  $V[\vec{Z} + \vec{a}] = V[\vec{Z}]$ .
- If  $\vec{a}_{k \times 1}$  is a non-random vector,  $V[\vec{a}^t \vec{Z}] = \vec{a}^t V[\vec{Z}] \vec{a}$ .
- If  $\mathbf{B}_{m \times k}$  is a non-random matrix,  $V[\mathbf{B}\vec{Z}] = \mathbf{B} V[\vec{Z}] \mathbf{B}^t$ .

The variance-covariance matrix of the sum of two random vectors has a simple property when the random vectors are independent:

- If  $\vec{Z}_{k \times 1}$  and  $\vec{U}_{k \times 1}$  are independent random vectors, then  $V[\vec{Z} + \vec{U}] = V[\vec{Z}] + V[\vec{U}]$ .

# The Multivariate Normal Distribution

Random vectors have multivariate probability distributions. The most important multivariate distribution is the **Multinormal**:

## Multivariate Normal Distribution

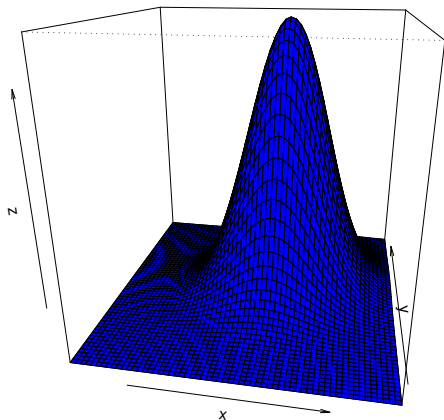
The  $k$ -dimensional random vector  $\vec{\mathbf{Z}}$  has a **Multinormal distribution**, with **parameters** given by the vector  $\vec{\boldsymbol{\mu}}$  and the invertible matrix  $\boldsymbol{\Sigma}$  if its joint density function is:

$$f(\vec{\mathbf{Z}}) = \frac{1}{(2\pi)^{k/2} \sqrt{\det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\vec{\mathbf{z}} - \vec{\boldsymbol{\mu}})^t \boldsymbol{\Sigma}^{-1} (\vec{\mathbf{z}} - \vec{\boldsymbol{\mu}})}, \quad \vec{\mathbf{z}} \in \mathbb{R}^k.$$

Notation:  $\vec{\mathbf{Z}} \sim \mathcal{N}_k(\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ .

**Warning:** A generalized Multinormal is also defined when matrix  $\boldsymbol{\Sigma}$  is not invertible, using the **generalized inverse**  $\boldsymbol{\Sigma}^-$ .

# The Binormal (Multinormal with $k = 2$ ) density function





# Some properties of the Multinormal distribution

## Theorem (Properties of Multinormal distribution)

If  $\vec{Z} \sim \mathcal{N}_k(\vec{\mu}, \Sigma)$ :

- 1 The expected vector of  $\vec{Z}$  is  $E[\vec{Z}] = \vec{\mu}$ .
- 2 The (co)variance matrix of  $\vec{Z}$  is  $V[\vec{Z}] = \Sigma$ .
- 3 If two components of  $\vec{Z}$  have zero covariance, they are independent:  
 $Cov(Z_i, Z_j) = 0 \Rightarrow Z_i, Z_j$  independent.

Note: In introductory Statistics courses it is shown that  $X, Y$  independent  $\Rightarrow cov(X, Y) = 0$ . When the joint distribution of  $X$  and  $Y$  is Multinormal, the converse implication is also true.

Note: Any zero in a (co)variance matrix of a Multinormal indicates that the corresponding components are independent.

# Properties of the Multinormal (cont.)

## Theorem (Properties of the Multinormal)

If  $\vec{Z} \sim \mathcal{N}_k(\vec{\mu}, \Sigma)$ :

- 4 All the marginal distributions of  $\vec{Z}$  are (multi)normal. In particular, each component  $Z_i$  is Normal with mean  $\mu_i$  and variance  $\Sigma_{(i,i)}$ :  
 $Z_i \sim \mathcal{N}(\mu_i, \Sigma_{(i,i)})$ .
- 5 If  $\vec{a}$  is a (non-random)  $k \times 1$  vector, then  $\vec{Z} + \vec{a} \sim \mathcal{N}_k(\vec{\mu} + \vec{a}, \Sigma)$ .
- 6 Linear combinations of the components of a Multinormal vector have Normal distributions:  $\vec{a}^t \vec{Z} = a_1 Z_1 + a_2 Z_2 + \dots + a_k Z_k \sim \mathcal{N}(\vec{a}^t \vec{\mu}, \vec{a}^t \Sigma \vec{a})$ .
- 7 If  $\mathbf{B}$  is a non-random matrix  $m \times k$  (of rank  $m \leq k$ ), then  $\mathbf{B}\vec{Z} \sim \mathcal{N}_m(\mathbf{B}\vec{\mu}, \mathbf{B}\Sigma\mathbf{B}^t)$ .

**Note:** In the latter result, if  $\mathbf{B}$  is a non-random matrix of rank  $m > k$ , the distribution of  $\mathbf{B}\vec{Z}$  has a **generalized** Multinormal distribution.

# Linear Regression Model - vector version

## The Linear Model in vector notation

$$1 \quad \vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}.$$

$$2 \quad \vec{\varepsilon} \sim \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n), \text{ with } \vec{\mathbf{0}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} ; \quad \sigma^2 \mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

In the second assumption, four statements are being made (considering the properties of Multinormal distributions discussed above):

- Each individual random error  $\varepsilon_i$  has a Normal distribution.
- Each individual random error has mean zero:  $E[\varepsilon_i] = 0$ .
- Each individual random error has the same variance:  $V[\varepsilon_i] = \sigma^2$ .
- Different random errors are independent, because  $Cov[\varepsilon_i, \varepsilon_j] = 0$  when  $i \neq j$  and, for a Multinormal, that implies independence.

# The distribution of $\vec{Y}$

The following Theorem is a direct consequence of slides 129 and 130.

## Theorem (First implications of the Model)

Given the Linear Regression Model, we have:

$$\vec{Y} \sim \mathcal{N}_n(\mathbf{X}\vec{\beta}, \sigma^2 \mathbf{I}_n).$$

In fact,  $\vec{Y}$  is the sum of the non-random vector ( $\mathbf{X}\vec{\beta}$ ) with the random vector ( $\vec{\epsilon}$ ):

$$\vec{Y} = \underbrace{\mathbf{X}\vec{\beta}}_{= "a''} + \underbrace{\vec{\epsilon}}_{= "z''}.$$

- $\vec{\epsilon} \sim \mathcal{N}(\vec{0}, \sigma^2 \mathbf{I}_n)$ .
- Adding a non-random vector ( $\mathbf{X}\vec{\beta}$ ) to a Multinormal random vector ( $\vec{\epsilon}$ ) does not destroy Multinormality.
- $E[\vec{Y}] = E[\mathbf{X}\vec{\beta} + \vec{\epsilon}] = \mathbf{X}\vec{\beta} + E[\vec{\epsilon}] = \mathbf{X}\vec{\beta}$ .
- $V[\vec{Y}] = V[\mathbf{X}\vec{\beta} + \vec{\epsilon}] = V[\vec{\epsilon}] = \sigma^2 \mathbf{I}_n$ .

# The distribution of $\vec{Y}$ (interpretation)

$$\vec{Y} \sim \mathcal{N}_n(\mathbf{X}\vec{\beta}, \sigma^2 \mathbf{I}_n).$$

Taking into account the properties of a Multinormal:

- Each individual observation  $Y_i$  has a Normal distribution.
- Each individual observation  $Y_i$  has mean value  $\mu_i = E[Y_i] = \vec{\mathbf{x}}_{[i]}^t \vec{\beta} = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \dots + \beta_p x_{p(i)}$ .
- Each individual observation has the same variance:  $V[Y_i] = \sigma^2$ .
- Different observations of  $Y$  are independent, because  $\text{Cov}[Y_i, Y_j] = 0$  when  $i \neq j$  and, in a Multinormal, that implies independence.

# The estimator of the Model parameters

We saw that vector  $\vec{\hat{\beta}}$  that estimates the vector  $\vec{\beta}$  of population parameters is:

$$\vec{\hat{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{Y},$$

where  $\mathbf{X}$  and  $\vec{Y}$  are the matrix and vector defined on slide 121.

Vector  $\vec{\hat{\beta}}$  has size  $p + 1$ . Its first element is the estimator of  $\beta_0$ , its second element is the estimator of  $\beta_1$ , etc...

In general, the estimator of  $\beta_j$  is in position  $j + 1$  of vector  $\vec{\hat{\beta}}$ .

The general results discussed above make it easy to determine the probability distribution of estimator  $\vec{\hat{\beta}}$ .

# The distribution of the vector of estimators $\vec{\hat{\beta}}$

## Theorem (Distribution of the estimator $\vec{\hat{\beta}}$ )

Given the Linear Regression Model, we have:

$$\vec{\hat{\beta}} \sim \mathcal{N}_{p+1}(\vec{\beta}, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}).$$

$\vec{\hat{\beta}}$  is the product of a non-random matrix,  $(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$ , and a random vector,  $\vec{\mathbf{Y}}$ :

$$\vec{\hat{\beta}} = \underbrace{(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t}_{\text{"B''}} \underbrace{\vec{\mathbf{Y}}}_{\text{"Z''}}.$$

- $\vec{\mathbf{Y}} \sim \mathcal{N}_n(\mathbf{X}\vec{\beta}, \sigma^2 \mathbf{I}_n)$ .
- Multiplying a non-random matrix,  $(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$ , by a Multinormal random vector ( $\vec{\mathbf{Y}}$ ) does not destroy **Multinormality**.
- $E[\vec{\hat{\beta}}] = E[(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{\mathbf{Y}}] = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t E[\vec{\mathbf{Y}}] = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{X} \vec{\beta} = \mathbf{I}_n \vec{\beta} = \vec{\beta}$ .
- $V[\vec{\hat{\beta}}] = V[(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{\mathbf{Y}}] = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t V[\vec{\mathbf{Y}}][(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t]^t = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \cdot \sigma^2 \mathbf{I}_n \cdot \mathbf{X}[(\mathbf{X}^t \mathbf{X})^{-1}]^t = \sigma^2 \cdot (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{X}[(\mathbf{X}^t \mathbf{X})^{-1}]^t = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}$ .

# The distribution of $\vec{\hat{\beta}}$ (interpretation)

$$\vec{\hat{\beta}} \sim \mathcal{N}_{p+1}(\vec{\beta}, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}).$$

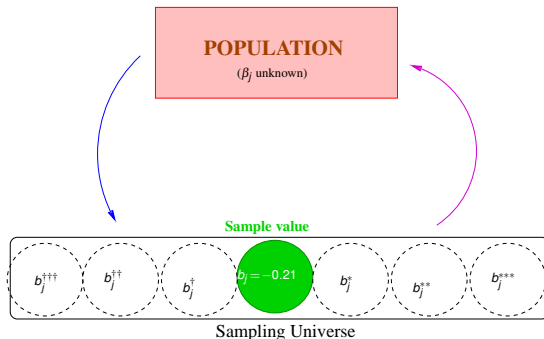
Taking into account the properties of a Multinormal (slides 129 and 130):

- Each individual estimator  $\hat{\beta}_j$  has a **Normal** distribution.
- Each individual estimator has mean value  $E[\hat{\beta}_j] = \beta_j$ , and is therefore **unbiased**.
- Each individual estimator has variance  $V[\hat{\beta}_j] = \sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1,j+1)}^{-1}$ .  
(Note the '+1' in the indices).
- Different individual estimators **are not** (in general) independent, because  $(\mathbf{X}^t \mathbf{X})^{-1}$  is not, in general, a diagonal matrix:  $Cov[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2 (\mathbf{X}^t \mathbf{X})_{(i+1,j+1)}^{-1}$ .
- Hence, the estimator  $\hat{\beta}_j$  of an individual parameter  $\beta_j$  has distribution  $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma_{\hat{\beta}_j}^2)$ , with  $\sigma_{\hat{\beta}_j}^2 = \sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1,j+1)}^{-1}$ .



# The sampling distribution of $\hat{\beta}_j$ (interpretation)

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma_{\hat{\beta}_j}^2) \quad \text{with} \quad \sigma_{\hat{\beta}_j}^2 = \sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1} .$$

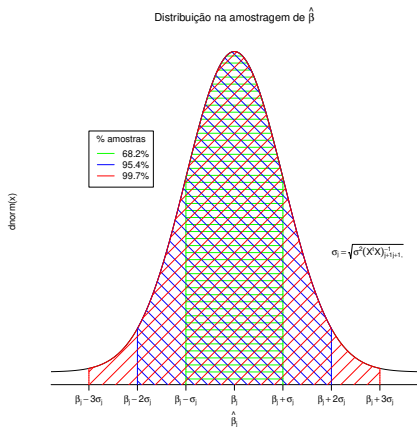


The set of all possible samples of size  $n$  is called the **Sampling Universe**.

The probability distribution of  $\hat{\beta}_j$  can be seen as the distribution of the values of  $b_j$  along the Sampling Universe.

# The sampling distribution of $\hat{\beta}_j$ (interpretation)

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma_{\hat{\beta}_j}^2) \quad \text{with} \quad \sigma_{\hat{\beta}_j}^2 = \sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1}.$$



# The distribution of an individual estimator

As was seen,  $\forall j = 0, 1, \dots, p$ :

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1}\right)$$
$$\Leftrightarrow \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim \mathcal{N}(0, 1),$$

with  $\sigma_{\hat{\beta}_j} = \sqrt{\sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1}}$ .

This distributional result would enable building confidence intervals or carrying out hypothesis tests on the parameters  $\vec{\beta}$ , were it not for the fact that the variance  $\sigma^2$  of the random errors is unknown.

# The problem of the unknown value of $\sigma^2$

In order to use the estimator  $\hat{\beta}_j$  for inference, we need to know its probability distribution, with no unknown quantities, other than  $\beta_j$ .

To overcome this problem, it is necessary to:

- find an estimator for  $\sigma^2$ ; and
- see what happens to the distribution of  $\hat{\beta}_j$  when  $\sigma^2$  is replaced by its estimator.

As  $\sigma^2 = V(\varepsilon_i)$ ,  $\forall i$ , and since the random errors  $\varepsilon_i$  are unknown, it is natural to seek an estimator of  $\sigma^2$  using the residuals.

# Estimating $\sigma^2$

Random errors (random variables – unobservable)

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \dots + \beta_p x_{p(i)})$$

Residuals (random variables – observable)

$$E_i = Y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \hat{\beta}_2 x_{2(i)} + \dots + \hat{\beta}_p x_{p(i)})}_{=\hat{Y}_i}$$

Residuals (observed)

$$e_i = y_i - (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \dots + b_p x_{p(i)})$$

The Maximum Likelihood estimator of  $\sigma^2$  (variance of the random errors) is:

$$\hat{\sigma}_{ML}^2 = \frac{SQRE}{n} .$$

But the estimator  $\hat{\sigma}_{ML}^2$  is biased:  $E \left[ \hat{\sigma}_{ML}^2 \right] = E \left[ \frac{SQRE}{n} \right] = \frac{n-(p+1)}{n} \cdot \sigma^2$

# The Residual Mean Square

A simple modification of the maximum likelihood estimator produces an unbiased estimator.

## Residual Mean Square (QMRE)

Define the **Residual Mean Square** as:

$$QMRE = \frac{SQRE}{n - (p + 1)} = \frac{\sum_{i=1}^n E_i^2}{n - (p + 1)}$$

Given a Linear Model,  $\hat{\sigma}^2 = QMRE$  is an unbiased estimator of the variance that is common to all random errors,  $\sigma^2 = V[\varepsilon_j]$ :

$$E[QMRE] = \sigma^2 .$$

The Residual Mean Square has as units of measurement the square of the units of  $Y$ .

# Pivots for inference on $\beta_j$

## Theorem (Distributions for inference on $\beta_j$ )

Given the Multiple Linear Regression Model, we have:

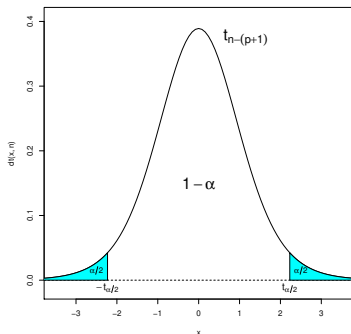
$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(p+1)}, \quad \forall j=0, 1, \dots, p$$

with  $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1}}$ .

This Theorem provides results that are at the root of building **confidence intervals** and **hypothesis tests** for the population parameters  $\beta_j$ .

## Deduction of confidence intervals for $\beta_j$

We know that  $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(p+1)}$ . Thus,



$$P \left[ -t_{\frac{\alpha}{2}} < \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} < t_{\frac{\alpha}{2}} \right] = 1 - \alpha$$



## Deduction CI for $\beta_j$ (cont.)

Work on the double inequality so as to isolate  $\beta_j$ :

$$P \left[ -t_{\frac{\alpha}{2}} < \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} < t_{\frac{\alpha}{2}} \right] = 1 - \alpha$$

$$\begin{aligned} & -t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j} < \hat{\beta}_j - \beta_j < t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j} \\ \Leftrightarrow & \quad t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j} > \beta_j - \hat{\beta}_j > -t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j} \\ \Leftrightarrow & \hat{\beta}_j - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j} < \beta_j < \hat{\beta}_j + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j}. \end{aligned}$$

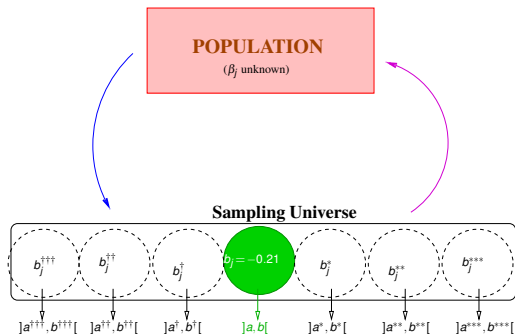
The **random interval**

$$\left] \hat{\beta}_j - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j} \right[$$

contains  $\beta_j$  with probability  $1 - \alpha$ .

# Random interval for $\beta_j$ (interpretation)

$$\left] \hat{\beta}_j - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j} \right[$$



Each sample in the **Sampling Universe** generates a **concrete interval**, called **Confidence Interval**.

A proportion  $1 - \alpha$  of those intervals contain the true value of  $\beta_j$ . The remaining  $\alpha$  do not contain  $\beta_j$ .

# Confidence Interval for $\beta_j$

## $(1 - \alpha) \times 100\%$ Confidence Interval for $\beta_j$

Given the Multiple Linear Regression Model and a sample, the  $(1 - \alpha) \times 100\%$  confidence interval for parameter  $\beta_j$  is:

$$\left[ b_j - t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j}, \quad b_j + t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j} \right],$$

with:

- $b_j$  element  $j+1$  of the vector of estimates  $\vec{\mathbf{b}}$  (slide 83);
- $t_{\frac{\alpha}{2}[n-(p+1)]}$  the quantile of order  $1 - \frac{\alpha}{2}$  in a  $t_{n-(p+1)}$  distribution;
- $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t \mathbf{X})_{(j+1,j+1)}^{-1}}$  (with the value of QMRE from our sample).

NOTE: The size of the CI increases with  $QMRE$  and the diagonal element of matrix  $(\mathbf{X}^t \mathbf{X})^{-1}$  associated with parameter  $\beta_j$ .

## Confidence Intervals for $\beta_j$ in

The information needed for computing the confidence intervals for each  $\beta_j$  can be obtained with the command `summary`. In the example on slide 101:

```
> summary(iris2.lm)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.24031    0.17837  -1.347    0.18
Petal.Length  0.52408    0.02449  21.399 < 2e-16 ***
Sepal.Length -0.20727    0.04751  -4.363 2.41e-05 ***
Sepal.Width   0.22283    0.04894   4.553 1.10e-05 ***
```

It is estimated that on average, the petal width decreases 0.20727 cm for each additional 1 cm in the sepal length (with other measurements fixed).

Since  $t_{0.025(146)} = 1.976346$ , the 95% CI for  $\beta_2$  is

$$\begin{aligned} & ] (-0.20727) - (1.976346)(0.04751) , (-0.20727) + (1.976346)(0.04751) [ \\ & \Leftrightarrow ] -0.3012 , -0.1134 [ \end{aligned}$$

## Confidence Intervals for $\beta_j$ in $\mathbb{R}$ (cont.)

Alternatively, it is possible to use the command `confint` to obtain the confidence intervals for each individual  $\beta_j$ :

```
> confint(iris2.lm)                                     <- 95% confidence (by default)
              2.5 %           97.5 %
(Intercept) -0.5928277      0.1122129
Petal.Length 0.4756798      0.5724865
Sepal.Length -0.3011547     -0.1133775
Sepal.Width  0.1261101      0.3195470

> confint(iris2.lm , level=0.99)                       <- 99% confidence
              0.5 %           99.5 %
(Intercept) -0.70583864     0.22522386
Petal.Length 0.46016260     0.58800363
Sepal.Length -0.33125352    -0.08327863
Sepal.Width  0.09510404     0.35055304
```

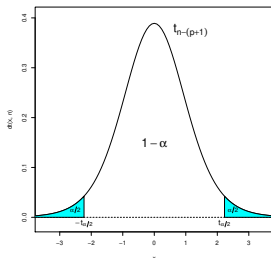
# Hypothesis Tests on the parameters

The result used to build CIs also enables us to carry out Hypothesis Tests on any  $\beta_j$ . Assuming the **Null Hypothesis**  $H_0 : \beta_j = c$ :

$$T = \frac{\hat{\beta}_j - \overbrace{\beta_j}^{=c} |_{H_0}}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(p+1)}, \quad \forall j=0, 1, \dots, p$$

We reject  $H_0$  for the **Alternative Hypothesis**  $H_1 : \beta_j \neq c$  if the computed value of  $T$  in the sample,  $T_{calc}$ , falls in one of the tails of the distribution.

Setting the **Significance Level**  $\alpha$ , we have the **Critical Region**:



# Hypothesis Test (bilateral) on $\hat{\beta}_j$

## Hypothesis Tests for $\beta_j$ (Multiple Linear Regression Model)

Hypotheses:  $H_0 : \beta_j = c$  vs.  $H_1 : \beta_j \neq c$

Test Statistic:  $T = \frac{\hat{\beta}_j - \overbrace{\beta_j}^{=c}|_{H_0}}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(p+1)}$  , if  $H_0$  is true.

Significance Level:  $\alpha$

Critical Region (bilateral Rejection Region): **Reject  $H_0$  when**

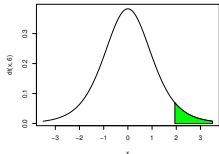
$$T_{calc} > t_{\frac{\alpha}{2}} [n-(p+1)] \quad \text{or} \quad T_{calc} < -t_{\frac{\alpha}{2}} [n-(p+1)]$$

$$\iff |T_{calc}| > t_{\frac{\alpha}{2}} [n-(p+1)]$$

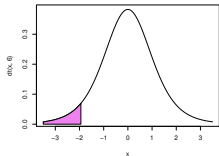
# Hypothesis Test on $\hat{\beta}_j$ (one-sided)

$$T = \frac{\hat{\beta}_j - \overbrace{\beta_{j|H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(p+1)}$$

With the **Alternative Hypothesis**  $H_1 : \beta_j > c$ , only large values of the test statistic suggest the rejection of  $H_0 : \beta_j \leq c$  (or  $H_0 : \beta_j = c$ ):



With the **Alternative Hypothesis**  $H_1 : \beta_j < c$ , only small values of  $T_{calc}$  suggest the rejection of  $H_0 : \beta_j \geq c$  (or  $H_0 : \beta_j = c$ ):





# Hypothesis Tests for the parameters

Given the Multiple Linear Regression Model,

## Hypothesis Tests for $\beta_j$ (Multiple Linear Regression)

Hypotheses:  $H_0 : \beta_j \begin{matrix} \geq \\ = \\ \leq \end{matrix} c$  vs.  $H_1 : \beta_j \begin{matrix} < \\ \neq \\ > \end{matrix} c$

Test Statistic:  $T = \frac{\hat{\beta}_j - \overbrace{\beta_j}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(p+1)}$ , if  $H_0$  is true.

Significance Level:  $\alpha$

Critical Region (Rejection Region): **Reject  $H_0$  when**

$$T_{calc} < -t_{\alpha[n-(p+1)]} \quad (\text{Left tail region})$$

$$|T_{calc}| > t_{\alpha/2[n-(p+1)]} \quad (\text{Two-tail region})$$

$$T_{calc} > t_{\alpha[n-(p+1)]} \quad (\text{Right tail region})$$

# Linear combinations of parameters

Let  $\vec{a} = (a_0, a_1, \dots, a_p)^t$  be a non-random vector in  $\mathbb{R}^{p+1}$ . The inner product  $\vec{a}^t \vec{\beta}$  defines a linear combination of the model parameters:

$$\vec{a}^t \vec{\beta} = a_0 \beta_0 + a_1 \beta_1 + a_2 \beta_2 + \dots + a_p \beta_p .$$

Important specific instances are when:

- $\vec{a}$  has a single non-zero element,  $a_{j+1} = 1$ :  $\vec{a}^t \vec{\beta} = \beta_j$ .
- $\vec{a}$  has only two non-zero elements,  $a_{i+1} = 1$  and  $a_{j+1} = \pm 1$ :  $\vec{a}^t \vec{\beta} = \beta_i \pm \beta_j$ .
- $\vec{a} = (1, x_1, x_2, \dots, x_p)$ :  $\vec{a}^t \vec{\beta}$  is the expected value of  $Y$  associated with the values indicated for the predictors:

$$\begin{aligned} \vec{a}^t \vec{\beta} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \\ &= E[Y | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p] \\ &= \mu_{Y|\vec{x}} \end{aligned}$$

## Inference on linear combinations of the $\beta_j$ s

$\vec{a}^t \vec{\beta}$  is estimated by the same linear combination of the estimators:

$$\vec{a}^t \vec{\hat{\beta}} = a_0 \hat{\beta}_0 + a_1 \hat{\beta}_1 + a_2 \hat{\beta}_2 + \dots + a_p \hat{\beta}_p .$$

We know the probability distribution of  $\vec{a}^t \vec{\hat{\beta}}$ :

- We know that  $\vec{\hat{\beta}} \sim \mathcal{N}_{p+1}(\vec{\beta}, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1})$  (slide 135);
- Hence,  $\vec{a}^t \vec{\hat{\beta}} \sim \mathcal{N}_1(\vec{a}^t \vec{\beta}, \sigma^2 \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a})$  (slide 130);
- That is,  $\vec{Z} = \frac{\vec{a}^t \vec{\hat{\beta}} - \vec{a}^t \vec{\beta}}{\sqrt{\sigma^2 \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}} \sim \mathcal{N}(0, 1)$ ;
- By a similar reasoning to that used when dealing with individual  $\beta_j$ , we have:

$$\frac{\vec{a}^t \vec{\hat{\beta}} - \vec{a}^t \vec{\beta}}{\sqrt{QMRE \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}} \sim t_{n-(p+1)} .$$

## Pivotal quantities for inference on $\vec{a}^t \vec{\beta}$

**Theorem** (A result for inference on linear combinations of  $\beta$ s)

Given the Multiple Linear Regression Model, we have

$$\frac{\vec{a}^t \hat{\vec{\beta}} - \vec{a}^t \vec{\beta}}{\hat{\sigma}_{\vec{a}^t \vec{\beta}}} \sim t_{n-(p+1)},$$

$$\text{com } \hat{\sigma}_{\vec{a}^t \vec{\beta}} = \sqrt{QMRE \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}.$$

This is a result that can be used to build **confidence intervals** and **hypothesis tests** for any linear combination of the model parameters  $\beta_j$ .

## Confidence Interval for $\vec{a}^t \vec{\beta}$

The similar structure to that of the pivotal quantity on slide 156 generates confidence intervals with a similar structure to those for individual  $\beta_j$ s.

### $(1 - \alpha) \times 100\%$ Confidence Interval for $\vec{a}^t \vec{\beta}$

Given the Linear Regression Model and a sample, the  $(1 - \alpha) \times 100\%$  confidence interval for a linear combination of the parameters,

$\vec{a}^t \vec{\beta} = a_0 \beta_0 + a_1 \beta_1 + \dots + a_p \beta_p$ , is:

$$\left[ \vec{a}^t \vec{b} - t_{\frac{\alpha}{2}, [n-(p+1)]} \cdot \hat{\sigma}_{\vec{a}^t \vec{\beta}}, \vec{a}^t \vec{b} + t_{\frac{\alpha}{2}, [n-(p+1)]} \cdot \hat{\sigma}_{\vec{a}^t \vec{\beta}} \right],$$

with  $\vec{a}^t \vec{b} = a_0 b_0 + a_1 b_1 + \dots + a_p b_p$  and  $\hat{\sigma}_{\vec{a}^t \vec{\beta}} = \sqrt{QMRE \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}$ .

## Formulas for inference on $\beta_i \pm \beta_j$

The general formula  $\hat{\sigma}_{\vec{\mathbf{a}}^t \vec{\hat{\beta}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{\mathbf{a}}}$  has an **alternative expression** in the specific instance of a **sum or difference of two  $\beta$ s**.

From the general formula for the variance of the sum or difference of random variables,

$$\begin{aligned} V[\hat{\beta}_i \pm \hat{\beta}_j] &= V[\hat{\beta}_i] + V[\hat{\beta}_j] \pm 2 \text{Cov}[\hat{\beta}_i, \hat{\beta}_j] . \\ \Leftrightarrow \sigma_{\hat{\beta}_i \pm \hat{\beta}_j}^2 &= \sigma^2 \cdot [(\mathbf{X}^t \mathbf{X})_{[i+1, i+1]}^{-1} + (\mathbf{X}^t \mathbf{X})_{[j+1, j+1]}^{-1} \pm 2 (\mathbf{X}^t \mathbf{X})_{[i+1, j+1]}^{-1}] . \end{aligned}$$

Hence, the standard error of  $\hat{\beta}_i \pm \hat{\beta}_j$  is:

$$\hat{\sigma}_{\hat{\beta}_i \pm \hat{\beta}_j} = \sqrt{QMRE \cdot [(\mathbf{X}^t \mathbf{X})_{[i+1, i+1]}^{-1} + (\mathbf{X}^t \mathbf{X})_{[j+1, j+1]}^{-1} \pm 2 (\mathbf{X}^t \mathbf{X})_{[i+1, j+1]}^{-1}] .}$$

## CI's for linear combinations in

In a Multiple Linear Regression, the confidence interval of a **generic linear combination**  $\vec{a}^t \vec{\beta}$ , requires the estimated (co)variance matrix of the estimators  $\vec{\hat{\beta}}$ ,

$$\widehat{V[\vec{\hat{\beta}}]} = QMRE \cdot (\mathbf{X}^t \mathbf{X})^{-1} .$$

This is given by the R command `vcov`.

The estimated (co)variance matrix in the **MLR** iris example is:

```
> vcov(iris2.lm)
              (Intercept)  Petal.Length  Sepal.Length  Sepal.Width
(Intercept)  0.031815766   0.0015144174  -0.005075942  -0.002486105
Petal.Length 0.001514417   0.0005998259  -0.001065046  0.000802941
Sepal.Length -0.005075942  -0.0010650465  0.002256837  -0.001344002
Sepal.Width  -0.002486105   0.0008029410  -0.001344002  0.002394932
```

## CI's for linear combinations in $\mathbb{R}$ (cont.)

The (estimated) standard error of  $\hat{\beta}_2 + \hat{\beta}_3$  (formula on slide 158) is:

$$\hat{\sigma}_{\hat{\beta}_2 + \hat{\beta}_3} = \sqrt{\hat{V}[\hat{\beta}_2 + \hat{\beta}_3]} = \sqrt{\hat{V}[\hat{\beta}_2] + \hat{V}[\hat{\beta}_3] + 2\hat{Cov}[\hat{\beta}_2, \hat{\beta}_3]}$$

$$\hat{\sigma}_{\hat{\beta}_2 + \hat{\beta}_3} = \sqrt{0.002256837 + 0.002394932 + 2(-0.001344002)} = 0.04431439.$$

```
> vcov(iris2.lm)
```

	(Intercept)	Petal.Length	Sepal.Length	Sepal.Width
(Intercept)	0.031815766	0.0015144174	-0.005075942	-0.002486105
Petal.Length	0.001514417	0.0005998259	-0.001065046	0.000802941
Sepal.Length	-0.005075942	-0.0010650465	0.002256837	-0.001344002
Sepal.Width	-0.002486105	0.0008029410	-0.001344002	0.002394932



# Confidence intervals for $E[Y|X_1 = x_1, \dots, X_p = x_p]$

Another specific case of the general result is of interest:

## CI for the expected value of $Y$ , given the predictor values

Given the Linear Regression Model and a sample with the values  $\vec{x} = (x_1, x_2, \dots, x_p)^t$  for the predictors, the expected value of  $Y$ ,

$$\mu_{Y|\vec{x}} = E[Y|X_1 = x_1, \dots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p ,$$

is estimated by  $\hat{\mu}_{Y|\vec{x}} = b_0 + b_1 x_1 + \dots + b_p x_p$  .

A  $(1 - \alpha) \times 100\%$  confidence interval for  $\mu_{Y|\vec{x}}$  is given by:

$$\left[ \hat{\mu}_{Y|\vec{x}} - t_{\frac{\alpha}{2}, [n-(p+1)]} \cdot \hat{\sigma}_{\hat{\mu}_{Y|\vec{x}}} , \hat{\mu}_{Y|\vec{x}} + t_{\frac{\alpha}{2}, [n-(p+1)]} \cdot \hat{\sigma}_{\hat{\mu}_{Y|\vec{x}}} \right] ,$$

with  $\hat{\sigma}_{\hat{\mu}_{Y|\vec{x}}} = \sqrt{QMRE \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}$ , where  $\vec{a} = (1, x_1, x_2, \dots, x_p)$ .

# Tests on linear combinations of parameters

Given the Linear Regression Model,

## Hypothesis Tests for $\vec{a}^t \vec{\beta}$

Hypotheses:  $H_0 : \vec{a}^t \vec{\beta} \begin{matrix} \geq \\ = \\ \leq \end{matrix} c$  vs.  $H_1 : \vec{a}^t \vec{\beta} \begin{matrix} < \\ \neq \\ > \end{matrix} c$

Test Statistic:  $T = \frac{\overbrace{\vec{a}^t \vec{\beta} - \vec{a}^t \vec{\beta}|_{H_0}}{=c}}{\hat{\sigma}_{\vec{a}^t \vec{\beta}}} \sim t_{n-(p+1)}$ , if  $H_0$  is true

Significance Level:  $\alpha$

Critical Region (Rejection Region): **Reject  $H_0$  when**

$$T_{calc} < -t_{\alpha[n-(p+1)]} \quad (\text{Left-tailed})$$

$$|T_{calc}| > t_{\alpha/2[n-(p+1)]} \quad (\text{Two-tailed})$$

$$T_{calc} > t_{\alpha[n-(p+1)]} \quad (\text{Right-tailed})$$

## Inference on $\mu_{Y|\vec{x}} = E[Y|\vec{x}]$ in

Estimated values and confidence intervals for  $\mu_{Y|\vec{x}}$  can be obtained with the command `predict`. The new predictor values are given in a data frame (with names equal to those in the original fit).

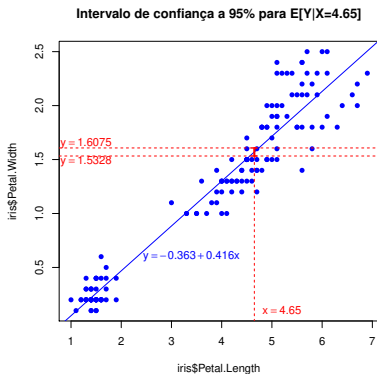
In the **Simple Linear Regression** iris example, the expected petal widths for flowers with petal lengths 1.85 and 4.65, are:

```
> predict(iris.lm, new=data.frame(Petal.Length=c(1.85,4.65)))  
      1      2  
0.406072 1.570187
```

## Inference for $E[Y|\vec{x}]$ in $\mathbb{R}$ (cont.)

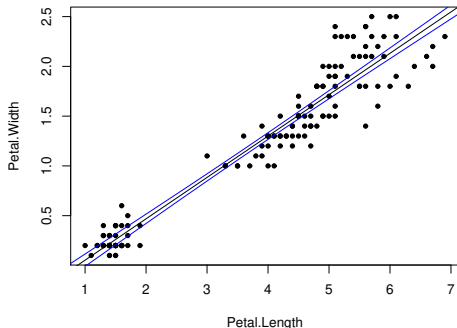
A confidence interval for  $\mu_{Y|\vec{x}}$  is obtained adding the argument `int="conf"`:

```
> predict(iris.lm,data.frame(Petal.Length=c(4.65)),int="conf")
      fit      lwr      upr
1 1.570187 1.5328338 1.6075405
```



# Confidence bands for the regression line

Considering the CIs for many values of  $x$  in some interval, we obtain a **confidence band** that contains the regression line with  $(1 - \alpha) \times 100\%$  confidence.



The confidence intervals for  $\mu_{Y|x}$  depend on the value of  $x$  (formula in slide 166). They will be **wider the further  $x$  is from the mean  $\bar{x}$  of the predictor observations**. Thus, the bands are **curved**.

# Formulas for a simple linear regression

In a simple linear regression, a formula for the variance  $\hat{\mu}_{Y|X}$  is:

$$\begin{aligned}\sigma_{\hat{\mu}_{Y|X}}^2 &= V[\hat{\mu}_{Y|X}] = \sigma^2 \cdot \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) \cdot s_X^2} \right] \\ \Rightarrow \hat{\sigma}_{\hat{\mu}_{Y|X}}^2 &= QMRE \cdot \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) \cdot s_X^2} \right].\end{aligned}$$

The confidence interval for  $\mu_{Y|X}$  in a Simple Linear Regression is:

$$\left] (b_0 + b_1 x) - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\mu}_{Y|X}} \ , \ (b_0 + b_1 x) + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\mu}_{Y|X}} \left[ .\right.$$

## MLR: Confidence intervals for $E[Y|\vec{x}]$ in

The command `predict` also enables us to obtain confidence intervals for  $\mu_{Y|\vec{x}}$  in a multiple linear regression.

In the multiple linear regression for the iris data, here is the 95% CI for the expected petal width for flowers with:

```
Petal.Length=2      Sepal.Length=5      Sepal.Width=3.1
```

```
> predict(iris2.lm, new=data.frame(Petal.Length=c(2),  
+   Sepal.Length=c(5), Sepal.Width=c(3.1)), int="conf")
```

```
      fit      lwr      upr  
[1,] 0.462297 0.4169203 0.5076736
```

The CI for  $E[Y | X_1 = 2, X_2 = 5, X_3 = 3.1]$  is: ] 0.4169203 , 0.5076736 [.

It is not possible to visualize this interval in  $\mathbb{R}^4$ .

# Variability of an individual observation of $Y$

We considered confidence intervals for the expected value of  $Y$ ,

$$\mu_{Y|\vec{x}} = E[Y|X_1=x_1, X_2=x_2, \dots, X_p=x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p,$$

which use the variability corresponding to estimator  $\hat{\mu}_{Y|\vec{x}}$ :

$$\sigma_{\hat{\mu}_{Y|\vec{x}}}^2 = V[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p] = \sigma^2 \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a},$$

with  $\vec{a} = (1, x_1, x_2, \dots, x_p)$ .

An individual observation of  $Y$  has additional variability, because:

$$Y = \mu_{Y|\vec{x}} + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon.$$

The random fluctuation of individual observations around the hyperplane is  $V[\varepsilon] = \sigma^2$ . It will be necessary to add the variance associated with the estimation of the hyperplane and the variance of individual observations:

$$\sigma_{Indiv}^2 = V[\hat{\mu}_{Y|\vec{x}}] + V[\varepsilon] = \sigma^2 \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a} + \sigma^2 = \sigma^2 \cdot [\vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a} + 1].$$



# Prediction intervals for $Y$

We can obtain **prediction intervals for individual observations of  $Y$** , associated with the predictor values  $X_1 = x_1, \dots, X_p = x_p$ .

In these intervals, the estimated variance of an individual observation of  $Y$  is the **estimate  $\sigma_{indiv}^2$** , that results by replacing  $\sigma^2$  with the sample value of **QMRE**:

## Prediction intervals for individual observations

$$\left[ \hat{\mu}_{Y|\bar{x}} - t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \quad , \quad \hat{\mu}_{Y|\bar{x}} + t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \right]$$

where

$$\hat{\mu}_{Y|X} = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p$$

and

$$\hat{\sigma}_{indiv} = \sqrt{QMRE [1 + \bar{\mathbf{a}}^t (\mathbf{X}^t \mathbf{X})^{-1} \bar{\mathbf{a}}]} \quad \text{com} \quad \bar{\mathbf{a}} = (1, x_1, x_2, \dots, x_p).$$

# Formulas for simple linear regressions

In a simple linear regression we can use the formula on slide 166:

$$\sigma_{Indiv}^2 = \underbrace{\sigma^2 \cdot \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) \cdot s_x^2} \right]}_{=V[\hat{\mu}_{Y|\bar{x}}]} + \underbrace{\sigma^2}_{=V[\varepsilon]} = \sigma^2 \cdot \left[ 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) \cdot s_x^2} \right].$$

Hence,

Simple LR: Prediction interval for an individual observation of Y

$$\left[ \hat{\mu}_{Y|x} - t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{Indiv} \quad , \quad \hat{\mu}_{Y|x} + t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{Indiv} \right].$$

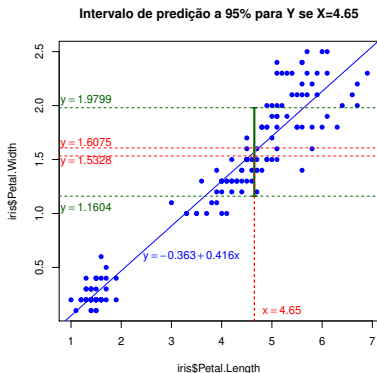
$$\text{com } \hat{\mu}_{Y|x} = b_0 + b_1 x \quad \text{e} \quad \hat{\sigma}_{Indiv} = \sqrt{QMRE \cdot \left[ 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) \cdot s_x^2} \right]}.$$

Both in simple and multiple linear regressions, these intervals are necessarily wider than the confidence intervals for  $\mu_{Y|\bar{x}}$  (for any given confidence level  $(1 - \alpha) \times 100\%$ ).

# Prediction intervals for $Y$ in

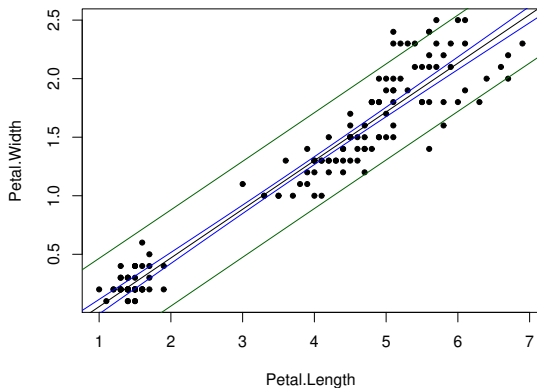
With R, a **prediction interval** for an individual observation of  $Y$  is obtained with the argument `int="pred"` in command `predict`:

```
> predict(iris.lm,data.frame(Petal.Length=c(4.65)), int="pred")  
      fit      lwr      upr  
1 1.570187 1.160442632 1.9799317
```



## Prediction bands for observations of $Y$

As with confidence intervals for  $E[Y|X = x]$ , varying the  $x$  values gives rise to **prediction bands** for individual values of  $Y$ .



## Prediction intervals for $Y$ (cont.)

With the iris multiple linear regression, the prediction interval for the petal width of an iris flower with petal length 2, and sepals of length 5 and width 3.1 is:

```
> predict(iris2.lm, data.frame(Petal.Length=c(2),  
+   Sepal.Length=c(5), Sepal.Width=c(3.1)), int="pred")
```

```
           fit           lwr           upr  
[1,] 0.462297 0.08019972 0.8443942
```

The requested prediction interval is: ] 0.0802 , 0.8444 [.

The corresponding confidence interval for  $\mu_{Y|\bar{x}}$  was ] 0.4169 , 0.5077 [, necessarily shorter.

# Testing the overall goodness of fit

In a **Linear Regression**, the model is **useless** if it is indistinguishable from the **Null Model**, i.e., the model with equation  $Y_i = \beta_0 + \varepsilon_i$ . The Null Model can be seen as a **submodel** of any linear model, in which **all** the predictors have coefficient zero:  $\beta_j = 0, \forall j > 0$ .

The **goodness-of-fit test** tests whether a given linear model is significantly different from the null model.

The hypotheses from which to choose are:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

[FULL MODEL  $\equiv$  NULL MODEL]

vs.

$$H_1 : \exists j = 1, \dots, p \text{ t.q. } \beta_j \neq 0$$

[FULL MODEL  $\neq$  NULL MODEL]

Note:  $\beta_0$  plays no role in the hypotheses.

## The goodness-of-fit test (cont.)

Defining:

- The **Regression Mean Square** as  $QMR = \frac{SQR}{p}$ .
- The **Residual Mean Square** as  $QMRE = \frac{SQRE}{n-(p+1)}$ .

If the goodness-of-fit Null Hypothesis is true:

$$F = \frac{QMR}{QMRE} \sim F_{[p, n-(p+1)]}.$$

This is the **F statistic** for the goodness-of-fit test.

## Alternative expression for the $F$ -test statistic

The statistic of the goodness-of-fit  $F$ -test in a Multiple Linear Regression model has an equivalent alternative expression:

$$F = \frac{n - (p + 1)}{p} \cdot \frac{R^2}{1 - R^2} .$$

The  $F$  statistic is an increasing function of the sample Coefficient of Determination,  $R^2$ . This justifies the right-hand sided Critical Region.

The test hypotheses can also be written as

$$H_0 : R^2 = 0 \quad \text{vs.} \quad H_1 : R^2 > 0 .$$

The hypothesis  $H_0 : R^2 = 0$  indicates the lack of a linear relation between  $Y$  and the predictors. It corresponds to a “disastrous” model fit. But its rejection is not synonymous with a good fit.



# The goodness-of-fit $F$ -test

## Goodness-of-fit $F$ -test with a Multiple Linear Regression

Hypotheses:  $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$

vs.

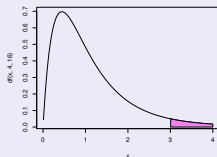
$H_1 : \exists j = 1, \dots, p$  such that  $\beta_j \neq 0$ .

Test statistic:  $F = \frac{QMR}{QMRE} \sim F_{[p, n-(p+1)]}$  if  $H_0$ .

Significance level:  $\alpha$

Crítical Region (Refection Region): One-sided, right-hand region

Reject  $H_0$  when  $F_{calc} > f_{\alpha[p, n-(p+1)]}$



# A different formulation of the goodness-of-fit $F$ -test

## $F$ -test for a Multiple Linear Regression (alternative)

Hypotheses:  $H_0 : \mathcal{R}^2 = 0$  vs.  $H_1 : \mathcal{R}^2 > 0$ .

Test statistic:  $F = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} \sim F_{[p, n-(p+1)]}$  if  $H_0$ .

Significance level:  $\alpha$

Critical Region (Rejection Region): One-sided (right)

Reject  $H_0$  when  $F_{calc} > f_{\alpha(p, n-(p+1))}$

The Null Hypothesis  $H_0 : \mathcal{R}^2 = 0$  states that, in the population, the coefficient of determination is zero.

# MLR inference example: the Brix data (Exercise 9)

Multiple Linear Regression of *Brix* over all other variables:

```
> brix.lm <- lm(Brix ~ . , data=brix)           <- note the use of '.'
> summary(brix.lm)
[...]
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	6.08878	1.00252	6.073	0.000298	***
Diametro	1.27093	0.51219	2.481	0.038030	*
Altura	-0.70967	0.41098	-1.727	0.122478	
Peso	-0.20453	0.14096	-1.451	0.184841	
pH	0.51557	0.33733	1.528	0.164942	
Acucar	0.08971	0.03611	2.484	0.037866	*

--

Residual standard error: 0.1366 on 8 degrees of freedom  
Multiple R-squared: 0.8483, Adjusted R-squared: 0.7534  
F-statistic: 8.944 on 5 and 8 DF, p-value: 0.003942

The final output line has the information for a goodness-of-fit  $F$  test.

The last 2 columns of table *Coefficients* provide information for the (bilateral)  $t$ -tests for each  $H_0 : \beta_j = 0$ .

# The principle of parsimony in MLR

Recall the **principle of parsimony** in modelling: we want a model that suitably describes the relation between the variables, but which is **as simple (parsimonious) as possible**.

If a Multiple Linear Regression model has a fit considered suitable, this principle suggests exploring whether **it is possible to find a submodel, with fewer predictors, without a significant loss of goodness-of-fit**.

# Model and Submodels

Given a Multiple Linear Regression model, with equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 ,$$

we call any linear regression with only some predictors a **submodel**.

E.g.,

$$Y = \beta_0 + \beta_2 x_2 + \beta_5 x_5 ,$$

The submodel can be identified by the set  $\mathcal{S}$  of predictors belonging to the submodel. In the example,  $\mathcal{S} = \{2, 5\}$ .

The model and submodel are identical if  $\beta_j = 0$  for any predictor  $x_j$  whose subscript does **not** belong to  $\mathcal{S}$ .

# Comparing a model with a submodel

To assess whether a given model significantly differs from one of its submodels (identified by the set  $\mathcal{S}$  of indices of its predictors), we must choose between the following hypotheses:

$$H_0 : \beta_j = 0, \quad \forall j \notin \mathcal{S} \quad \text{vs.} \quad H_1 : \exists j \notin \mathcal{S} \quad \text{such that} \quad \beta_j \neq 0.$$

[SUBMODEL OK]

[SUBMODEL WORSE]

NOTE: This discussion only involves coefficients  $\beta_j$  of predictor variables. The intercept  $\beta_0$  is always part of the submodel equations.

The intercept  $\beta_0$  is irrelevant from the point of view of parsimony: it does not require additional work when collecting the data, nor in interpreting the model. But it ensures better fits.

## A test statistic for model/submodel comparison

Consider a **full model** with  $p$  predictors and Residual Sum of Squares  $SQRE_C$ ; and a **submodel** with  $k$  predictors and Residual Sum of Squares  $SQRE_S$

Given the Null Hypothesis:

$\beta_j = 0$  for all variables  $x_j$  that do not belong to the submodel,

we have:

$$F = \frac{\frac{SQRE_S - SQRE_C}{p-k}}{\frac{SQRE_C}{n-(p+1)}} \sim F_{[p-k, n-(p+1)]},$$

**Note:** The denominator  $\frac{SQRE_C}{n-(p+1)}$  is the Residual Mean Square of the full model,  $QMRE_C$ .

# The test for a submodel (partial $F$ test)

## $F$ -test comparing a model with one of its submodels

Given the Multiple Linear Regression Model,

Hypotheses:

$$H_0 : \beta_j = 0, \quad \forall j \notin \mathcal{S} \quad \text{vs.} \quad H_1 : \exists j \notin \mathcal{S} \quad \text{such that} \quad \beta_j \neq 0.$$

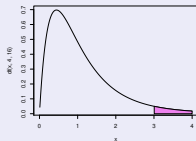
Test statistic:

$$F = \frac{(SQRE_S - SQRE_C)/(p-k)}{SQRE_C/[n-(p+1)]} \sim F_{[p-k, n-(p+1)]}, \text{ sob } H_0.$$

Significance level:  $\alpha$

Critical Region (Rejection Region): One-sided, right

$$\text{Reject } H_0 \text{ if } F_{calc} > f_{\alpha[p-k, n-(p+1)]}$$





## Alternative expression for the test statistic

The test statistic  $F$  to compare a full model with  $p$  predictors and one of its submodels with only  $k$  predictors can alternatively be written as:

$$F = \frac{n - (p + 1)}{p - k} \cdot \frac{R_C^2 - R_S^2}{1 - R_C^2}.$$

The test hypotheses can also be written as:

$$H_0 : R_C^2 = R_S^2 \quad \text{vs.} \quad H_1 : R_C^2 > R_S^2,$$

The hypothesis  $H_0$  indicates that the strength of the linear relation between  $Y$  and the set of predictors is identical in the model and in the submodel.

If we do not reject  $H_0$ , we choose the submodel (more parsimonious).

If we reject  $H_0$ , we choose the full model (with a significantly better fit).

# Partial $F$ test: alternative formulation

## Partial $F$ -test to compare a model with one of its submodels

Given the Multiple Linear Regression Model,

Hypotheses:

$$H_0 : R_C^2 = R_S^2 \quad \text{vs.} \quad H_1 : R_C^2 > R_S^2 .$$

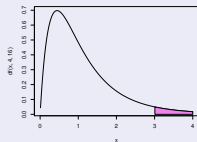
Test statistic:

$$F = \frac{n-(p+1)}{p-k} \cdot \frac{R_C^2 - R_S^2}{1 - R_C^2} \sim F_{[p-k, n-(p+1)]}, \text{ under } H_0 .$$

Significance level:  $\alpha$

Critical Region (Rejection Region): One-sided, right region

Reject  $H_0$  when  $F_{calc} > f_{\alpha[p-k, n-(p+1)]}$



# Testing submodels with

The necessary information for a partial  $F$  test is obtained with the command `anova`, with **two arguments**: the `lm` object resulting from fitting the **full model** and the **submodel** with which it is being compared.

With the `iris` dataset examples, we have:

```
> anova(iris.lm, iris2.lm)
```

```
Analysis of Variance Table
```

```
Model 1: Petal.Width ~ Petal.Length
```

```
Model 2: Petal.Width ~ Petal.Length + Sepal.Length + Sepal.Width
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	148	6.3101				
2	146	5.3803	2	0.9298	12.616	8.836e-06 ***

The values  $R_s^2 = 0.9271$  and  $R_c^2 = 0.9379$  of the models `iris.lm` and `iris2.lm` are **significantly different**.

## Partial $F$ -test relations

The **goodness-of-fit test** is equivalent to a partial  $F$  test comparing a linear model and its Null submodel (with no predictors).

If the model and submodel differ by a single predictor,  $X_j$ , the partial  $F$  test is equivalent to the  $t$ -test (slide 153) with the hypotheses  $H_0 : \beta_j = 0$  vs.  $H_1 : \beta_j \neq 0$ .

In this case, not only are the the hypotheses of both tests the same, as the test statistic for the partial  $F$  test is the square of the associated  $t$ -test statistic.

In a **simple** linear regression, the  $t$ -test on a zero slope is equivalent to the goodness-of-fit  $F$  test. The latter test statistic is the square of the former.

## (No) How to choose a submodel?

The partial  $F$  test (for nested models) allows us to choose between a model and a submodel. Sometimes a submodel is suggested by:

- **theoretical reasons**, which suggest that certain predictors may not, in fact, be important in predicting the values of  $Y$ .
- **practical reasons**, such as the difficulty, cost or workload associated with collecting observations or setting up an experiment with certain predictors.

In such cases, it may be clear which submodels are to be tested.

## (No) How to choose a submodel? (cont.)

But in many situations it is not initially clear which subsets of predictors are to be retained in the submodel. The only aim is to see whether the model may be simplified. In such cases, choosing a submodel is not an easy problem.

Given  $p$  predictors, the number of possible subsets, with any number of predictors, except 0 (the empty set) and  $p$  (the full model) that can be chosen is  $2^p - 2$ . The following table indicates the number of such subsets for  $p = 5, 10, 15, 20$ .

$p$	$2^p - 2$
5	30
10	1 022
15	32 766
20	1 048 574

## (No) Beware of simultaneous exclusion of predictors

For small values of  $p$ , it is possible to analyse all possible subsets. With appropriate algorithms and computer software, a full search of all possible subsets is still possible for larger values of  $p$  (up to  $p \approx 35$ ). But for very large values of  $p$ , a full search is computationally unfeasible.

We cannot justify the **joint** exclusion of several predictors based on the  $t$ -tests for the significance of each single coefficient  $\beta_j$  in the full model.

In fact, the  $t$ -tests on each coefficient  $\beta_j$  assume that all the remaining predictors are in the model. The exclusion of any predictor changes the model fit: it changes the estimated values  $b_j$ , and their standard errors, for the predictors that remain in the submodel. It may happen that a predictor can be dropped from a full model, but not from a submodel, or vice-versa.

## (No) An example: the Brix data (Exercise 9)

The **individual** exclusion of three predictors is admissible (for  $\alpha = 0.05$ ):

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	6.08878	1.00252	6.073	0.000298	***
Diametro	1.27093	0.51219	2.481	0.038030	*
Altura	-0.70967	0.41098	-1.727	0.122478	
Peso	-0.20453	0.14096	-1.451	0.184841	
pH	0.51557	0.33733	1.528	0.164942	
Acucar	0.08971	0.03611	2.484	0.037866	*

But it is **not** legitimate to claim that the **joint** exclusion of *Altura*, *Peso* and *pH* will not significantly affect the goodness-of-fit.

```
> anova(brix2.lm,brix.lm)
```

Analysis of Variance Table

Model 1: Brix ~ Diametro + Acucar

Model 2: Brix ~ Diametro + Altura + Peso + pH + Acucar

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	11	0.42743					
2	8	0.14925	3	0.27818	4.97	0.03104	*



## (No) Full searches

For a small or medium number  $p$  of predictors, there are algorithms and software routines that perform a **complete search** and determine the subset of  $k$  predictors with the largest value of  $R^2$  (or some other criterion of model quality).

The **leaps and bounds** algorithm of Furnival and Wilson <sup>2</sup> is a computationally efficient algorithm that identifies the best subset of predictors, for a given cardinality  $k$ .

A software implementation of the algorithm is available in R, in the **package leaps** (command with the same name). A similar routine can be found in command **e leaps** in the package **subselect**.

---

<sup>2</sup>Furnival, G.W and Wilson, R.W.,Jr. (1974) Regressions by leaps and bounds, *Technometrics*, **16**, 499-511.

## (No) Example using the leaps function

Despite the small number of predictors, we illustrate the use of the `leaps` command with the `brix` dataset.

```
> colnames(brix)      <-- to see the variable names
[1] "Diametro" "Altura"  "Peso"      "Brix"      "pH"        "Acucar"

> library(leaps)     <-- to load the package (must be installed)

> leaps(y=brix$Brix, x=brix[,-4], method="r2", nbest=1) <-- arguments: y response, x predictors
$which              <-- logical matrix, specifying the selected predictors
  1     2     3     4     5 <-- columns: predictors; rows: size k of subset
1 FALSE FALSE FALSE FALSE TRUE <-- k=1 ; best individual predictor: Acucar
2 TRUE  TRUE FALSE FALSE FALSE <-- k=2 ; best pair of predictors: Diametro and Altura
3 TRUE  TRUE FALSE FALSE TRUE  <-- k=3 ; best trio of predictors: Diametro, Altura and Acucar
4 TRUE  TRUE FALSE TRUE  TRUE
5 TRUE  TRUE TRUE  TRUE  TRUE
[...]
```

	1	2	3	4	5	
1	FALSE	FALSE	FALSE	FALSE	TRUE	<-- k=1 ; best individual predictor: Acucar
2	TRUE	TRUE	FALSE	FALSE	FALSE	<-- k=2 ; best pair of predictors: Diametro and Altura
3	TRUE	TRUE	FALSE	FALSE	TRUE	<-- k=3 ; best trio of predictors: Diametro, Altura and Acucar
4	TRUE	TRUE	FALSE	TRUE	TRUE	
5	TRUE	TRUE	TRUE	TRUE	TRUE	

```
$r2                <-- Coef. Determination of best solution with k=1,2,3,4,5 predictors
[1] 0.5091325 0.6639105 0.7863475 0.8083178 0.8482525
```

Notice how the best two-predictor submodel (highest  $R_s^2$ ) is **not** the submodel with the predictors `Diametro` and `Acucar`, as suggested by the  $p$ -values in the full model.


## (No) Stepwise search algorithms

Alternatively, computationally lighter **search algorithms** may be used, that **do not analyse all possible submodels and do not guarantee the identification of the best subsets**.

Simple algorithms of this kind are **sequential**, adding or dropping **one predictor at each step of the algorithm**, until some **stopping rule** is met. In particular, stepwise algorithms may be:

- **backward elimination** when, starting with the full model, the exclusion of a single variable is considered at each step of the algorithm.
- **forward selection** when, starting with the Null Model, the inclusion of one variable is considered at each step.
- **stepwise selection** when, for a given pre-specified direction, exclusions and inclusions are alternatively considered.

## (No) Stepwise algorithms based on the AIC

 provides functions that automate stepwise searches for submodels, in which the criterion to exclude/include a variable at each step is based on the **Akaike Information Criterion (AIC)**.

The AIC is a **general indicator of the goodness-of-fit of models** based in the Likelihood function. In the context of a **Linear Regression with  $k$  predictors**, it is defined as:

### Akaike Information Criterion in a Linear Model

$$AIC = n \cdot \ln \left( \frac{SQRE_k}{n} \right) + 2(k + 1).$$

AIC values of different Linear Models can be compared, **as long as they are fitted with the same dataset and have the same response variable  $Y$** .

## (No) Interpretation of the AIC

$$AIC = n \cdot \ln \left( \frac{SQRE_k}{n} \right) + 2(k+1).$$

- The **first term** measures the **goodness-of-fit** of the model to the dataset. **The smaller, the better.**
- The **second term** measures **model complexity**, through the number of predictors. **The smaller, the better.**

A model for the response variable  $Y$  is considered better than another if its **AIC is lower** (this favours models with smaller  $SQRE$ , but also with fewer predictors).

The **AIC can be used to select between a model and any of its submodels**. Submodels always have larger values of  $SQRE$ , but smaller values of  $k$ . Whether the **AIC value is smaller depends on the trade-off**.

## (No) Stepwise algorithms based on the AIC (cont.)

In a backward elimination algorithm, based on the AIC criterion:

- the full model is fitted and its AIC is computed.
- all possible submodels with one predictor less are fitted and their AICs computed.
- If none of the submodel AICs is smaller than the AIC of the current model, the algorithm stops and the running model is the final one. If dropping some variables reduces the AIC, we exclude the predictor for which the AIC drops the most and return to the previous point.

## (No) Stepwise search algorithms in

The command `step` runs a stepwise selection algorithm based on the AIC. Consider again the `brix` dataset example:

```
> step(brix.lm, dir="backward")
```

```
Start:  AIC=-51.58
```

```
Brix ~ Diametro + Altura + Peso + pH + Acucar
```

	Df	Sum of Sq	RSS	AIC
<none>			0.14925	-51.576
- Peso	1	0.039279	0.18853	-50.306
- pH	1	0.043581	0.19284	-49.990
- Altura	1	0.055631	0.20489	-49.141
- Diametro	1	0.114874	0.26413	-45.585
- Acucar	1	0.115132	0.26439	-45.572

In this case, **no predictor is excluded**: the AIC of the initial (full) model is smaller than that of any submodel resulting from dropping a single predictor. **The final model is the initial model.**

## (No) A final word on search algorithms

Stepwise selection algorithms do **not** guarantee the selection of the best submodel with a given number of predictors. They only identify, in a computationally “cheap” way, submodels that are “good”.

They should be used with common sense and the resulting submodels considered taking other aspects into account (for example, the cost or difficulty in collecting the data, or the role of each predictor in theoretical terms for the problem at hand).



## Model validation and other diagnostics

A Linear Regression analysis is not complete without a study of the residuals and other diagnostic tools.

The Linear Model assumes  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$  ,  $\forall i = 1, \dots, n$ . We cannot **directly** check these assumptions: random errors are unobservable.

### Distribution of Residuals, given the Model

Given the linear model, the **residuals** have the following distribution:

$$E_i \sim \mathcal{N}\left(0, \sigma^2(1 - h_{ii})\right) \quad \forall i = 1, \dots, n,$$

with  $h_{ij}$  the  $i$ -th diagonal element of the matrix  $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$  of orthogonal projections on the subspace  $\mathcal{C}(\mathbf{X})$ .

This result can be proved by considering the vector of residuals,

$$\vec{E} = \vec{Y} - \vec{\hat{Y}} = \vec{Y} - \mathbf{H}\vec{Y} = (\mathbf{I}_n - \mathbf{H})\vec{Y}.$$

# Properties of Residuals given the linear model

## Theorem (Distribution of Residuals with the Linear Model)

Given the Linear Model, we have:

$$\vec{\mathbf{E}} \sim \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2(\mathbf{I}_n - \mathbf{H})) \quad \text{sendo} \quad \vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}.$$

Since with the Linear Model  $\vec{\mathbf{Y}} \sim \mathcal{N}(\mathbf{X}\vec{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_n)$ , the vector of residuals  $\vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$ , has a generalized **Multinormal** distribution (slide 130).

The expected vector of  $\vec{\mathbf{E}}$  results from the properties of slide 125:

- $E[\vec{\mathbf{E}}] = E[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})E[\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\vec{\boldsymbol{\beta}} = \vec{\mathbf{0}}$ ,  
since  $\mathbf{X}\vec{\boldsymbol{\beta}} \in \mathcal{C}(\mathbf{X})$ , remaining invariant when projected:  $\mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}} = \mathbf{X}\vec{\boldsymbol{\beta}}$ .
- From the properties of slide 126 and since  $\mathbf{H}$  is **symmetric** ( $\mathbf{H}^t = \mathbf{H}$ ) and **idempotent** ( $\mathbf{H}\mathbf{H} = \mathbf{H}$ ), we have:  
 $V[\vec{\mathbf{E}}] = V[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})V[\vec{\mathbf{Y}}](\mathbf{I}_n - \mathbf{H})^t = \sigma^2 \cdot (\mathbf{I}_n - \mathbf{H})$ .

## Properties of Residuals in the Linear Model (cont.)

**Note:** Although in the Linear Model random errors are independent, **residuals are not independent random variables**: their covariances are not (in general) zero:

$$\text{cov}(E_i, E_j) = -\sigma^2 \cdot h_{ij}, \quad \text{se } i \neq j,$$

where  $h_{ij}$  is the element on row  $i$ , column  $j$  of matrix  $\mathbf{H}$ .

If  $\vec{\mathbf{E}} \sim \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$ , then individual residuals have distribution:

$$E_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii})),$$

where  $h_{ii}$  is the  $i$ -th diagonal element of  $\mathbf{H}$  and

$$\frac{E_i}{\sqrt{\sigma^2(1 - h_{ii})}} \sim \mathcal{N}(0, 1).$$

## Two types of residuals

Since  $\frac{E_i}{\sqrt{\sigma^2(1-h_{ii})}} \sim \mathcal{N}(0, 1)$ , standardized residuals are defined:

Usual residuals :  $E_i = Y_i - \hat{Y}_i$ ;

Standardized residuals :  $R_i = \frac{E_i}{\sqrt{QMRE \cdot (1-h_{ii})}}$ .

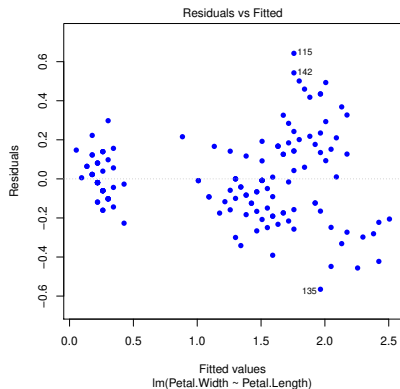
For large samples,  $R_i$  are approximately  $\mathcal{N}(0, 1)$ .

The R command `rstandard` calculates standardized residuals ( $R_i$ ).

In linear regressions, the validity of model assumptions is checked using residual plots. Normality tests are not carried out because residuals are not (in general) independent.

# Model checking: (1) scatterplots of residuals vs. $\hat{Y}_i$

A necessary scatterplot: (usual) Residuals vs. fitted values of  $Y$ .



- Residuals should be in a horizontal band around zero.
- without any apparent pattern: given the Linear Model,  $cor(E_i, \hat{Y}_i) = 0$ .

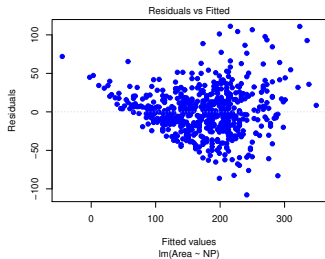
# Patterns suggesting problems

In scatterplots of  $E_i$  vs.  $\hat{Y}_i$  patterns may appear:

**Curvature:** Suggests violation of the assumed linearity between  $y$  and the predictors.

**Funnel-shaped pattern:** Suggests violation of the variance homogeneity assumption.

**One or more points strongly deviated from the trend:** Indicates the presence of outliers.



A **funnel-shaped** plot, also suggesting some **curvature** (videiras dataset, Exercise 18, Area vs. NP).

## Model checking: (2) Plots to assess Normality

As was seen on slide 204, for large samples the standardized residuals

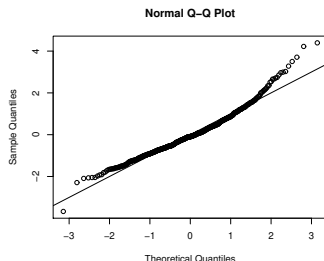
$$R_i = \frac{E_i}{\sqrt{QMRE \cdot (1 - h_{ii})}}, \text{ are approximately } \mathcal{N}(0, 1).$$

The assumption of Normal random errors may be validated with:

- a **qq-plot** comparing the **empirical quantiles** of the  $n$  standardized residuals, with the corresponding **theoretical quantiles** of a  $\mathcal{N}(0, 1)$ .

A qq-plot validates the Normality assumption if it is approximately collinear.

This qq-plot suggests some deviation from Normality:



## Model checking: (3) Plots to assess independence

Non-independence between random errors may result from:

- correlation along time;
- spatial correlation.

It may be useful to inspect plots of residuals vs. order of observation or the spatial distribution of residuals, to check for patterns suggesting lack of independence. If so, alternative time-series or spatial models may be needed.

## Model checking: (4) Plots of residuals vs. predictors

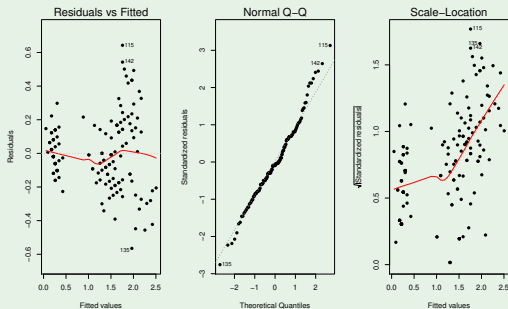
Non-linearity in plots of residuals vs. each individual predictor may suggest the need for transformation of those predictors.



# Model checking with

The command `plot`, when applied to an `lm` object produces up to six plots of residuals and other diagnostics. The first three are **residual plots**. For the `iris` example:

```
> plot(iris.lm, which=1:3, pch=16)
```



The third plot (argument `which=3`) is of  $\sqrt{|R_i|}$  vs.  $\hat{Y}_i$ .

## Other diagnostics

Other diagnostic tools seek to identify observations that deserve further scrutiny.

**Outliers** is a concept without a rigorous definition. It refers to observations that are far apart from the underlying linear trend between  $Y$  and the predictors.

Outliers are often associated with large (in absolute value) residuals. In particular, and since standardized residuals are approximately distributed as  $\mathcal{N}(0, 1)$  for large sample size  $n$ , observations for which  $|R_i| > 3$  may be classified as outliers.

But beware: observations very far from the underlying trend may have such an impact on the model fit that they no longer have large-magnitude residuals.

# Leverage points

The **leverage** of the  $i$ -th observation is defined as the  $i$ -th diagonal element of matrix  $\mathbf{H}$ :  $h_{ii} = \mathbf{H}_{(i,i)}$ .

Since  $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$ , we have  $\hat{y}_i = \sum_{j=1}^n h_{ij}y_j$  (each fitted value is a linear combination of the observed values). The leverage  $h_{ii}$  is the weight associated with  $y_i$  when defining the corresponding fitted value,  $\hat{y}_i$ . It should not be excessive.

**Leverage points** are observations with large  $h_{ii}$ . They tend to “attract” the fitted hypersurface.

Since  $V[E_i] = \sigma^2(1 - h_{ii})$ , if  $h_{ii}$  is large, the variance of the residual  $E_i$  is small and the residual will be close to its mean (zero). In other words, the fitted hypersurface tends to be close to that point.

## Leverage (cont.)

For **any** observation we have:

$$\frac{1}{n} \leq h_{ii} \leq 1 .$$

The **mean value** of the leverages in a linear regression is the ratio between the number of model parameters and the number of observation:

$$\bar{h} = \frac{p+1}{n} ,$$

Thus, **the more observations**, the smaller the mean leverage.

Observations with a high leverage **may, or may not, be outliers**.

## Leverage (cont.)

### Leverage in a Simple Linear Regression

In a **simple linear regression**, the leverage is given by the formula:

$$h_{ij} = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{(n-1) \cdot s_x^2}.$$

Thus, in a simple linear regression, the leverage of observation  $i$  depends on the distance of the predictor value  $x_i$  from the mean  $\bar{x}$ : the larger  $(x_i - \bar{x})^2$ , the larger the leverage  $h_{ij}$ . The largest leverage must belong to one of the two most extreme observations in  $x$ .

In a Multiple Linear Regression, large leverages also tend to be associated with the points whose predictor values are furthest from the vector of mean predictor values.

## Influential observations

**Influential observations** are observations that, if withdrawn from the dataset, produce noticeable changes in the fitted parameters  $b_j$  and fitted values  $\hat{y}_i$ .

The most frequent measure of **influence** is **Cook's distance**, defined as:

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{[-i]_j})^2}{(p+1) \cdot QMRE},$$

where  $\hat{y}_{[-i]_j}$  is the fitted value of observation  $i$ , that would result from fitting the  $\beta_j$ s **without observation  $i$** . An equivalent expression is:

$$D_i = R_i^2 \cdot \left( \frac{h_{ii}}{1 - h_{ii}} \right) \cdot \frac{1}{p+1}$$

The larger  $D_i$ , the greater the influence of the  $i$ -th observation.

A common rule is to consider  $D_i > 0.5$  as defining an influential observation.

# Diagnostic tools in

Command `hatvalues` computes leverages ( $h_{ij}$ ) and `cooks.distance` the  $D_i$ s.

## Brix dataset (Exercise 9)

```
> brix.diagn <- cbind(hatvalues(brix.lm), cooks.distance(brix.lm))
> colnames(brix.diagn) <- c("h_ii", "Di")
> brix.diagn
```

	h_ii	Di
1	0.6231274	0.6209707369
2	0.3576171	0.0969006496
3	0.4750339	0.0380279990
4	0.2881782	0.0186723249
5	0.3751686	0.0351359851
6	0.2985676	0.0354362871
7	0.5260699	0.0793008032
8	0.4955231	0.0304136309
9	0.2809899	0.2009993314
10	0.2268779	0.0002254622
11	0.2757540	0.0108143657
12	0.4771373	0.0092558438
13	0.6609377	1.5222084206
14	0.6390174	1.0769004225

Some very large values result from a small dataset ( $n=14$ ) for a heavy model ( $p=5$ ).

The mean leverage is  $\bar{h} = \frac{p+1}{n} = 0.4286$ .

# Warning

Outliers, influential observations and leverage points, although possibly related, are not the same concept.

For example, an observation with a large standardized residual and  $h_{ii}$  large, must also have a large Cook's distance, and therefore be influential. But if  $R_i^2$  is large and  $h_{ii}$  small (or vice versa), it may, or may not, be influential, depending on their relative values.

These diagnostics are useful essentially to identify observations that deserve greater attention.

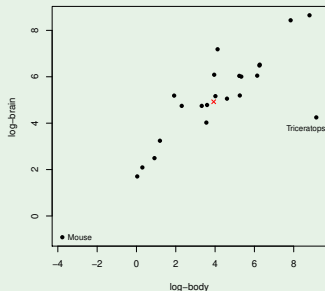


# A Simple Linear Regression example

## Animals data (Exercise 6)

Considering only a **subset** of the species, we obtain the following plot of log brain weight vs. log body weight:

```
> library(MASS)
> animaissub <- Animals[-c(6,19,25,26,27),]
> plot(log(brain) ~ log(body) , data=animaissub, pch=16)
```



## A Simple Linear Regression example (cont.)

Here are the resulting standardized residuals, Cook distances and leverages:

	R_i	D_i	h_ii	
Mountain beaver	-0.547	0.018	0.109	
Cow	-0.201	0.001	0.068	
Grey wolf	0.057	0.000	0.044	
Goat	0.168	0.001	0.045	
Guinea pig	-0.754	0.039	0.119	
Asian elephant	1.006	0.069	0.120	
Donkey	0.276	0.002	0.052	
Horse	0.121	0.001	0.071	
Potar monkey	0.711	0.015	0.057	
Cat	-0.006	0.000	0.081	
Giraffe	0.145	0.001	0.071	
Gorilla	0.195	0.001	0.053	
Human	1.850	0.078	0.044	
African elephant	0.688	0.046	0.163	
Triceratops	-3.610	1.431	0.180	<- D_i very large; h_ii not so much
Rhesus monkey	1.306	0.058	0.064	
Kangaroo	-0.578	0.008	0.044	
Mouse	-1.172	0.355	0.341	<- largest h_ii; D_i not so much
Rabbit	-0.519	0.013	0.089	
Sheep	0.163	0.001	0.044	
Jaguar	-0.243	0.001	0.046	
Chimpanzee	0.992	0.022	0.043	
Pig	-0.471	0.006	0.052	

## Diagnostic plots in

The command `plot`, when applied to an `lm` object produces, besides the plots considered in slide 209, also the following plots with other diagnostics:

Argument `which=4` produces a barplot of the Cook distance of each observation.

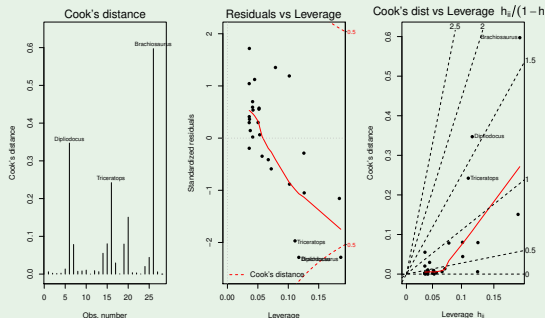
Argument `which=5` produces a scatterplot of the standardized residuals ( $R_i$ s) on the vertical axis against leverages  $h_{ij}$  on the horizontal axis, drawing isolines for some Cook distances (by default, 0.5 and 1), to highlight influential observations.

Argument `which=6` produces a scatterplot of Cook's distances (vertical axis) vs. values of  $\frac{h_{ij}}{1-h_{ij}}$ , with isolines for standardized residuals  $R_i$  (resulting from the formula on slide 214).

# An example of diagnostic plots

Here are these diagnostic plots for the `Animals` dataset (Ex. 6):

```
> plot(Animals.lm, which=4:6)
```



The large Cook distances reflect the dinosaurs' deviation from the general underlying trend for other species. The fact that there are **three** discordant observations somewhat reduces the value of those distances.

## The adjusted $R^2$

The usual Coefficient of Determination is defined as:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQRE}{SQT}$$

The **adjusted  $R^2$** , with  $QMT = \frac{SQT}{n-1} = s_y^2$ , is:

$$R_{mod}^2 = 1 - \frac{QMRE}{QMT} = 1 - \frac{SQRE}{SQT} \cdot \frac{n-1}{n-(p+1)} = 1 - (1 - R^2) \cdot \frac{n-1}{n-(p+1)}.$$

For any linear model (with predictors), we have:  $R_{mod}^2 < R^2$ .

If  $n \gg p+1$  (many more observations than parameters),  $R^2 \approx R_{mod}^2$ .

If  $n$  is little larger than  $p$ ,  $R_{mod}^2 \ll R^2$  (except when  $R^2 \approx 1$ ).

$\frac{QMRE}{QMT} = \frac{\hat{\sigma}^2}{s_y^2}$  measures the total variability of  $Y$  that remains unexplained when the predictors are introduced. Hence,  $R_{mod}^2$  is a measure of the gain in explaining variance  $s_y^2$  associated with the model.

## the adjusted $R^2$ (cont.)

The adjusted  $R_{mod}^2$  penalizes complex models that are fitted with few observations. Exercise 9: brix data ( $n=14$  and  $p+1=6$ ).

```
> summary(brix.lm)
[...]  
Multiple R-squared:  0.8483, Adjusted R-squared:  0.7534
```

The adjusted  $R^2$  of a submodel may be larger than that of a model.

### Example: Exercise 19

((No) also illustrates the use of  $R_{mod}^2$  as a selection criterion with the leaps function):

```
> library(leaps)
> leaps(y=milho$y , x=milho[,-10], method="adjr2", nbest=1)
[...]  
$adjr2      <-- the largest adjusted R2 is for the submodel with k=4 predictors  
[1] 0.5493014 0.6337329 0.6544835 0.6807418 0.6798986 0.6779395 0.6745412  
[8] 0.6633467 0.6488148
```

## (No) Some variable transformations

Sometimes, it is possible to overcome violations of the assumptions regarding the Normality or variance homogeneity of the random errors by **transforming variables**. For example,

$$\text{If } \text{var}(\varepsilon_i) \propto E[Y_i] \quad \text{then } Y \longrightarrow \sqrt{Y}$$

$$\text{If } \text{var}(\varepsilon_i) \propto (E[Y_i])^2 \quad \text{then } Y \longrightarrow \ln Y$$

$$\text{If } \text{var}(\varepsilon_i) \propto (E[Y_i])^4 \quad \text{then } Y \longrightarrow 1/Y$$

are standard proposals to stabilize variances.

The above examples are specific cases of the **Box-Cox family of transformations**:

$$Y \longrightarrow \begin{cases} \frac{Y^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ \ln(Y) & , \lambda = 0 \end{cases}$$

## (No) Warning about transformations

But the transformation of variables, especially when it affects the response variable, must be done with caution.

- A transformation of variables also changes the underlying trend relation between the original variables;
- A transformation that “corrects” one problem (e.g., variance heterogeneity) may create another (e.g., non-normality);
- There is the danger that using transformations that solve problems in a concrete sample may may not work in general.



## (No) Linearizing transformations

Different is the issue of transformations that seek to linearize a non-linear relation between a response variable and the predictors. Such **linearizing transformations** may also be useful with multiple linear regressions.

E.g., the non-linear relation between  $Y$ ,  $x_1$  and  $x_2$ ,

$$Y = \beta_0 x_1^{\beta_1} x_2^{\beta_2}$$

becomes, by taking logarithms, a linear relation between  $\ln(Y)$ ,  $\ln(x_1)$  and  $\ln(x_2)$  (with  $\beta_0^* = \ln(\beta_0)$ ):

$$\ln(Y) = \beta_0^* + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) .$$

## (No) Warnings about linearising transformations

- The estimates that minimise the sum of squared residuals in the linearised relations **are not** the same as the **optimal solutions to the problem of minimising the sum of squared residuals in the original non-linear relation**.
- The transformations discussed did not involve the random errors.
- The assumptions of additive, Normal, independent random errors with constant variance and zero mean **must be valid in the linear relations between the transformed variables**.

# Final warnings

1. Problems associated with (near) **multi-collinearity** of the predictors, that is, when the columns of the model matrix  $\mathbf{X}$  are (almost) linearly dependent:

- there may be numerical problems when calculating  $(\mathbf{X}^t\mathbf{X})^{-1}$ , thus in fitting the model and estimating the parameters;
- some  $\hat{\beta}_i$ s may have very large variances, resulting in very imprecise or unstable inference.

Multi-collinearity reflects redundancy of information in the predictors. It can be overcome by excluding from the dataset one or more predictors that are responsible for the (near) linear dependence of the predictors.

## Final warnings (cont.)

2. Do not confuse the existence of a linear relation between the predictors  $X_1, X_2, \dots, X_p$  and the response variable  $Y$ , with a cause and effect relation.

There **may** exist a cause-and-effect relation. But it may also be the case that there:

- a **joint variation**, but not of a causal nature (as for example, with many morphometric datasets). Sometimes, predictors and response variable are all reflecting common underlying causes.
- A **spurious** relation, with a numerical coincidence.

A **causal** relation can only be asserted based on some theory of the phenomenon under consideration, not on the statistically established linear relation.

# Analysis of Variance (ANOVA)

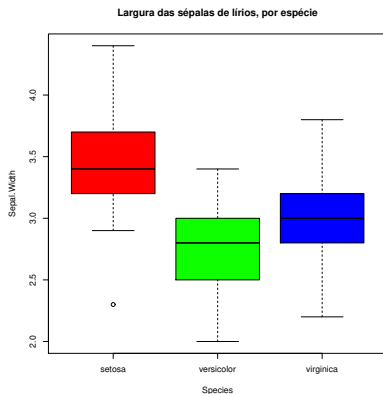
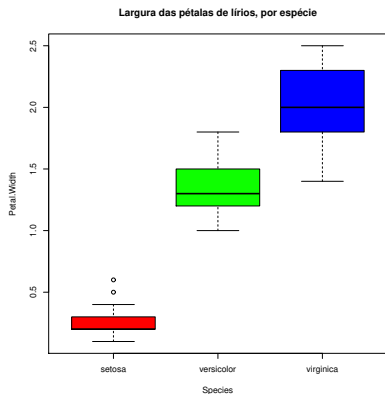
Linear Regressions model a numerical (quantitative) response variable, using one or more predictors, that are **also numerical**.

But a numerical response variable may be modelled with **qualitative (categorical)** variables, that is, one or more **factors**.

The **Analysis of Variance (ANOVA)** is a statistical methodology to deal with this type of situations.

ANOVAs were developed in the 1930s, in the Rothamstead Agricultural Experimental Station (England), by **R.A. Fisher**.

# Two examples: iris by species



Petal widths seem to differ between the iris species.  
Sepal lengths differ less.

Can the observed differences be attributed to real differences in the mean population values of each species?

# A ANOVA as a specific instance of the Linear Model

Although the Analysis of Variance arose as a separate method, both the Analysis of Variance and Linear Regressions are specific instances of the **Linear Model**.

Introducing ANOVA through its similarities with Linear Regression enables us to make use of much of the theory studied so far.

## Terminology:

**Response variable**  $Y$ : a **numerical** (quantitative) variable, that we wish to model.

**Factor** : a **categorical** (qualitative) predictor;

**Factor levels** : the different categories (“values”) of a factor, that is, different experimental situations where observations of  $Y$  are collected.

# One-way ANOVA

In a **one-way ANOVA**, the (numerical) response variable is modelled using a single categorical predictor (factor).

We assume we have  $n$  independent observations of the response variable  $Y$ , with  $n_i$  ( $i = 1, \dots, k$ ) corresponding to factor level  $i$ . Thus,

$$n_1 + n_2 + \dots + n_k = n.$$

## One-way balanced designs

When there is an equal number of observations from each factor level,

$$n_1 = n_2 = n_3 = \dots = n_k \quad (= n_c),$$

we speak of a **balanced design**.

For different reasons, **balanced designs** are advisable.



# Double indexation of $Y$

In regressions we indexed each of the  $n$  observations of  $Y$  with a single subscript, ranging from 1 to  $n$ .

In this new context, it is preferable to use **two indices to denote each observation of  $Y$** :

- one ( $i$ ) denotes the **factor level to which the observation corresponds**;
- the other ( $j$ ) allows the **identification of each observation within a given factor level**.

Thus, **the  $j$ -th observation of  $Y$ , in the  $i$ -th factor level**, is represented by  $Y_{ij}$ , (with  $i=1, \dots, k$  and  $j=1, \dots, n_i$ ).

## A model for $Y_{ij}$

we assume that the values of  $Y$  may differ because:

- they correspond to **different factor levels**; or
- due to **random (unexplained) variability**.

The poorer nature of our predictor implies a simpler model equation than in regressions.

In general, we assume that the **expected (mean) value of  $Y$  may differ in the  $k$  experimental situations (factor levels) in which it is observed.**

A first formulation of the model equation is:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{com} \quad E[\varepsilon_{ij}] = 0 .$$

Here,  $\mu_i$  represents the expected value of the observations  $Y_{ij}$ , collected in factor level  $i$ .

## A model for $Y_{ij}$ (cont.)

In order to fit ANOVA in the theory of Linear Models already studied, it is convenient to re-write the equation with a common additive constant:

$$E[Y_{ij}] = \mu_i = \mu + \alpha_j .$$

The parameter  $\mu$  will be common to all observations, while the parameters  $\alpha_j$  are specific to each factor level ( $i$ ).

Each  $\alpha_j$  is called the  $i$ -th level effect.

We assume that  $Y_{ij}$  randomly fluctuates around its mean value:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} ,$$

with  $E[\varepsilon_{ij}] = 0$ .

# The 1-way ANOVA Model as a Linear Model

The general equation

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} ,$$

means that:

- the  $n_1$  observations from level  $i = 1$  are modelled as  $Y_{1j} = \mu + \alpha_1 + \varepsilon_{1j}$ ;
- the  $n_2$  observations from level  $i = 2$  are  $Y_{2j} = \mu + \alpha_2 + \varepsilon_{2j}$ ;
- and so on...

In order to fit this set of equations within the context of the linear model, the general equation can be written as:

$$Y_{ij} = \mu + \alpha_1 \mathcal{I}_{1ij} + \alpha_2 \mathcal{I}_{2ij} + \dots + \alpha_k \mathcal{I}_{kij} + \varepsilon_{ij} ,$$

where the **dummy (indicator) variables** for each factor level are defined as:

$$\mathcal{I}_{mij} = \begin{cases} 1 & \text{se } i = m , \\ 0 & \text{se } i \neq m . \end{cases}$$

# The equation in vector notation

The model equation for a one-way ANOVA can be written in vector/matrix format, as in the linear regression model. Consider:

- $\vec{Y}$  the  $n$ -dimensional vector with all observations of the response variable. Assume that the  $n_1$  first correspond to factor level 1, the following  $n_2$  to level 2, and so on.
- $\vec{1}_n$  the vector of  $n$  ones, already considered in regression.
- $\vec{I}_i$  the vector of the indicator (dummy) variable for factor level  $i$ . For each observation, this variable takes the value 1 if the observation is from factor level  $i$ , and value 0 otherwise ( $i = 1, \dots, k$ ). In an ANOVA, the indicator variables play the role of the predictors.
- $\vec{\epsilon}$  the vector of  $n$  random errors.

# The vectors of the indicator variables

For example, if we have  $n = 9$  observations, with:

- $n_1 = 3$  observations from the first factor level;
- $n_2 = 4$  from the second level; and
- $n_3 = 2$  observations from the third level;

vectors  $\vec{\mathcal{I}}_2$  and  $\vec{\mathcal{I}}_3$  will be:

$$\vec{\mathcal{I}}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mathcal{I}}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

## The model equation in vector notation

In matrix/vector notations, the basic equation describing the  $n$  observations of  $Y$  may be written as in the Linear Model:

$$\begin{aligned}\vec{Y} &= \mu \vec{\mathbf{1}}_n + \alpha_1 \vec{\mathcal{J}}_1 + \alpha_2 \vec{\mathcal{J}}_2 + \alpha_3 \vec{\mathcal{J}}_3 + \vec{\boldsymbol{\varepsilon}} \\ \Leftrightarrow \vec{Y} &= \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}.\end{aligned}$$

The columns of matrix  $\mathbf{X}$  are the vector of  $n$  ones and the indicator variables. The vector of parameters  $\vec{\boldsymbol{\beta}}$  contains  $\mu$  and the level effects  $\alpha_j$ .

In the example with  $n_1 = 3$ ,  $n_2 = 4$  and  $n_3 = 2$  observations:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

# The problem of over-parametrization

There is a “technical” problem: the columns of such a matrix  $\mathbf{X}$  are linearly dependent, so that matrix  $\mathbf{X}^t\mathbf{X}$  is not invertible. There are too many parameters in the model. Possible solutions are:

- 1 drop the parameter  $\mu$  from the model.
  - ▶ this corresponds to dropping the column of ones from matrix  $\mathbf{X}$ ;
  - ▶ each  $\alpha_j$  becomes the factor level mean  $\mu_j$ ;
  - ▶ this solution cannot be generalized to more complex situations;
  - ▶ it is harder to fit into the Linear Model theory that we considered.
- 2 impose restrictions upon the parameters: e.g.,  $\sum_{j=1}^k \alpha_j = 0$ .
  - ▶ this is the classical solution, frequent in ANOVA literature;
  - ▶ it is harder to fit into the Linear Model theory.
- 3 impose the restriction  $\alpha_1 = 0$ : we will use this solution.
  - ▶ it drops the first indicator variable from the model (and from  $\mathbf{X}$ );
  - ▶ the Linear Model theory can be directly used, and the solution can be extended to more factors.

Each solution has implications in terms of the interpretation of parameters.



## The basic equation with our example

Assuming  $\alpha_1 = 0$ , we re-write the model equation as:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

Now  $\mu = \mu_1$  is the mean value of the observations from level  $i = 1$ :

$$\begin{aligned} E[Y_{1j}] &= \mu_1 & , \forall j = 1, \dots, n_1 \\ E[Y_{2j}] &= \mu_2 = \mu_1 + \alpha_2 & , \forall j = 1, \dots, n_2 \\ E[Y_{3j}] &= \mu_3 = \mu_1 + \alpha_3 & , \forall j = 1, \dots, n_3 \end{aligned}$$

## The level effects $\alpha_j$

In the one-way ANOVA model equation (slide 235), each  $\alpha_j$  ( $i > 1$ ) represents the **variation** that transforms the mean of level 1 into the mean of level  $i$ :

$$\alpha_1 = 0$$

$$\alpha_2 = \mu_2 - \mu_1$$

$$\alpha_3 = \mu_3 - \mu_1$$

$$\vdots \quad \vdots \quad \vdots$$

$$\alpha_k = \mu_k - \mu_1$$

The **equality of all population level means  $\mu_j$**  is equivalent to having **all level effects equal to zero:  $\alpha_j = 0$ ,  $\forall i$ .**

This is the **Null Hypothesis** in testing the existence of factor level effects.

# The one-way ANOVA model for inference

Adding the remaining assumptions of the Linear Model:

## One-way (1 factor) ANOVA model, with $k$ levels

There are  $n$  observations,  $Y_{ij}$ ,  $n_i$  of which correspond to factor level  $i$  ( $i = 1, \dots, k$ ). Assume:

- 1  $Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}$ ,  $\forall i=1, \dots, k$ ,  $\forall j=1, \dots, n_i$  (with  $\alpha_1 = 0$ ).
- 2  $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ ,  $\forall i, j$
- 3  $\{\varepsilon_{ij}\}_{i,j}$  independent random variables.

The model has  $k$  unknown parameters: the mean of  $Y$  in the first factor level,  $\mu_1$  and the effects  $\alpha_i$  ( $i > 1$ ) for each of the  $k-1$  remaining factor levels. In other words, the vector of parameters is:

$$\vec{\beta} = (\mu_1, \alpha_2, \alpha_3, \dots, \alpha_k)^t.$$

# The one-way ANOVA model - vector notation

Equivalently, in vector notation,

## One-way ANOVA model - vector notation

$$\textcircled{1} \quad \vec{Y} = \mu_1 \vec{\mathbf{1}}_n + \alpha_2 \vec{\mathcal{J}}_2 + \alpha_3 \vec{\mathcal{J}}_3 + \dots + \alpha_k \vec{\mathcal{J}}_k + \vec{\epsilon} = \mathbf{X}\vec{\beta} + \vec{\epsilon}, \quad \text{with}$$

- ▶  $\vec{Y}$  the random vector of the  $n$  observations of the response variable;
- ▶  $\vec{\mathbf{1}}_n$  the vector of  $n$  ones;
- ▶  $\vec{\mathcal{J}}_2, \vec{\mathcal{J}}_3, \dots, \vec{\mathcal{J}}_k$  the indicator variables for the stated levels;
- ▶  $\mathbf{X} = \left[ \vec{\mathbf{1}}_n \mid \vec{\mathcal{J}}_2 \mid \vec{\mathcal{J}}_3 \mid \dots \mid \vec{\mathcal{J}}_k \right]$  the model matrix; and
- ▶  $\vec{\beta} = (\mu_1, \alpha_2, \alpha_3, \dots, \alpha_k)^t$ .

$$\textcircled{2} \quad \vec{\epsilon} \sim \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n), \text{ with } \mathbf{I}_n \text{ the } n \times n \text{ identity matrix.}$$

It is a **Linear Model**, like the Multiple Linear Regression model, that only differs in the nature of the predictors, which here are the indicator variables for factor levels 2 to  $k$ .

# The test for factor effects

The hypothesis that no factor levels affect the mean of the response variable is the hypothesis

$$\begin{aligned} & \alpha_2 = \alpha_3 = \dots = \alpha_k = 0 \\ \Leftrightarrow & \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k \end{aligned}$$

Given the analogy with Linear Regression models, this hypothesis corresponds to stating that the coefficients of all the “predictor variables” (in this ANOVA, the dummy variables  $\vec{\mathcal{J}}_i$ ) are zero.

Thus, **this hypothesis can be tested using the model's goodness-of-fit  $F$  test** (slide 177).

In this context there are specific formulas.

# Degrees of freedom

In a one-way ANOVA, the number of predictors in the model (indicator variables for levels  $j > 1$ ) is  $p = k - 1$  and the number of model parameters is  $p + 1 = k$ .

We denote **SQF** (from **F**actor), instead of **SQR**, the Sum of Squares associated with the model fit.

The degrees of freedom associated with each Sum of Squares are:

SQxx	d.f.
SQF	$k - 1$
SQRE	$n - k$

The **Mean Squares** (QMF and QMRE) are the ratios of the Sums of Squares divided by their corresponding degrees of freedom.

# The $F$ test for factor effects in a one-way ANOVA

## $F$ test for factor effects

Given the one-way ANOVA Model, we have:

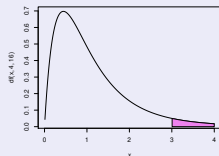
Hypotheses:  $H_0 : \alpha_i = 0 \quad \forall i=2, \dots, k$  vs.  $H_1 : \exists i=2, \dots, k \text{ t.q. } \alpha_i \neq 0$ .  
[NO FACTOR EFFECTS] vs. [FACTOR EFFECTS]

Test statistic:  $F = \frac{QMF}{QMRE} \sim F_{[k-1, n-k]}$  se  $H_0$ .

Significance level:  $\alpha$

Critical Region (Rejection Region): One-sided, right-tailed

Rej.  $H_0$  if  $F_{calc} > f_{\alpha[k-1, n-k]}$



Sums of Squares and Mean Squares have specific formulas in this context.

# Matrix $\mathbf{X}$ in a one-way ANOVA

Since the ANOVA model is a specific instance of the Linear Model, the formula for the least-squares estimators of the parameters is:

$$\vec{\hat{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}},$$

and the vector of fitted values  $\vec{\hat{\mathbf{Y}}}$  results from orthogonally projecting  $\vec{\mathbf{Y}}$  onto the subspace  $\mathcal{C}(\mathbf{X})$  of the columns of matrix  $\mathbf{X}$ :  $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$ .

But model matrix  $\mathbf{X}$  has a special nature: since its  $k$  columns are the vectors  $\vec{\mathbf{1}}_n, \vec{\mathcal{I}}_2, \vec{\mathcal{I}}_3, \dots, \vec{\mathcal{I}}_k$ , the elements of matrix  $\mathbf{X}$  in the ANOVA are all 0 or 1.

As a result, both the projection matrix  $\mathbf{H}$  and the vector  $\vec{\hat{\mathbf{Y}}}$ , have a specific nature.



# The fitted values $\hat{Y}_{ij}$

In a one-way ANOVA, any vector in the column-space  $\mathcal{C}(\mathbf{X})$  has equal values for all observations in the same factor level:

$$a_1 \vec{\mathbf{1}}_n + a_2 \vec{\mathcal{J}}_2 + a_3 \vec{\mathcal{J}}_3 + \dots + a_k \vec{\mathcal{J}}_k = \begin{bmatrix} a_1 \\ \dots \\ a_1 \\ \hline a_1 + a_2 \\ \dots \\ a_1 + a_2 \\ \hline a_1 + a_3 \\ \dots \\ a_1 + a_3 \\ \hline (\dots) \\ \hline a_1 + a_k \\ \dots \\ a_1 + a_k \end{bmatrix}$$

Vector  $\vec{\mathbf{Y}}$  belongs to  $\mathcal{C}(\mathbf{X})$ , and so has this nature.

## The fitted values $\hat{Y}_{ij}$

Specifically, in vector  $\vec{\hat{Y}} = \mathbf{H}\vec{Y}$ , all values  $\hat{Y}_{ij}$  for factor level  $i$  are given by the sample mean of the  $n_i$  observations  $Y_{ij}$  for that level:

$$\hat{Y}_{ij} = \bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij},$$

Note that to minimise the Residual Sum of Squares,

$$SQRE = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2,$$

and since all fitted values  $\hat{Y}_{ij}$  are the same for all observations in a common factor level  $i$ , we minimise the sum for each level by taking:  $\hat{Y}_{ij} = \bar{Y}_i$ .

# The fitted parameters

The population parameters are  $\mu_1$  and  $\alpha_j = \mu_j - \mu_1$ .

These population parameters are estimated by the corresponding sample quantities:

## Estimated parameters in a one-way ANOVA

$$\begin{aligned}\hat{\mu}_1 &= \bar{Y}_1. \\ \hat{\alpha}_2 &= \hat{\mu}_2 - \hat{\mu}_1 = \bar{Y}_2. - \bar{Y}_1. \\ \hat{\alpha}_3 &= \hat{\mu}_3 - \hat{\mu}_1 = \bar{Y}_3. - \bar{Y}_1. \\ &\vdots \quad \quad \quad \vdots \\ \hat{\alpha}_k &= \hat{\mu}_k - \hat{\mu}_1 = \bar{Y}_k. - \bar{Y}_1.\end{aligned}$$

## Residuals, SQRE and QMRE

We saw (slide 250) that  $\hat{Y}_{ij} = \hat{\mu}_i = \bar{Y}_{i.}$ , so that the residual of observation  $Y_{ij}$  is given by:

$$E_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i.},$$

Hence, the **Sum of Squared Residuals** is given by:

$$SQRE = \sum_{i=1}^k \sum_{j=1}^{n_i} E_{ij}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^k (n_i - 1) S_i^2,$$

where  $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$  is the sample variance of the  $n_i$  observations of  $Y$  in the  $i$ -th factor level. **SQRE** measures variability **within** the  $k$  levels.

The **Residual Mean Square** is a weighted mean of the level variances  $S_i^2$ , with weights  $n_i - 1$  ( $n - k = \sum_i (n_i - 1)$ ):

$$QMRE = \frac{SQRE}{n - k} = \frac{1}{n - k} \sum_{i=1}^k (n_i - 1) S_i^2.$$

# Formulas for balanced designs

In the case of a **balanced design**, i.e.,  $n_1 = n_2 = \dots = n_k (= n_c)$  and  $n = n_c \cdot k$ , and so:

$$SQRE = (n_c - 1) \sum_{i=1}^k s_i^2$$

$$QMRE = \frac{SQRE}{n - k} = \frac{n_c - 1}{n - k} \sum_{i=1}^k s_i^2 = \frac{n_c - 1}{k(n_c - 1)} \sum_{i=1}^k s_i^2 = \frac{1}{k} \sum_{i=1}^k s_i^2,$$

Thus, in balanced designs, the Residual Mean Square *QMRE* is the (simple) mean of the  $k$  level variances,  $s_i^2$ .

# The Factor Sum of Squares

Let  $\bar{Y}_{..} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$  be the overall mean of all  $n$  observations.

The **Factor Sum of Squares**,  $SQF$ , is given by:

$$\begin{aligned} SQF &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{Y}_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ \Leftrightarrow SQF &= \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \end{aligned}$$

$SQF$  measures the **variability among the sample means for each level**.

# Formulas for balanced designs

In the case of a **balanced design**,

$$SQF = n_c \sum_{i=1}^k (\bar{Y}_{i.} - \bar{Y}_{..})^2 = n_c(k-1) \cdot S_{Y_{i.}}^2,$$

where  $S_{Y_{i.}}^2 = \frac{1}{k-1} \sum_{i=1}^k (\bar{Y}_{i.} - \bar{Y}_{..})^2$  indicates the **sample variance** of the  $k$  level means in the sample.

$$QMF = \frac{SQF}{k-1} = n_c \cdot S_{Y_{i.}}^2.$$

Thus, in balanced designs, the Factor Mean Square,  $QMF$ , is proportional to the variance of the  $k$  level means of variable  $Y$ .

# The relation between Sums of Squares

The fundamental relation between the three Sums of Squares (even with unbalanced designs) has a **special meaning**:

$$\begin{aligned} SQT &= SQF + SQRE \\ \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 &= \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^k (n_i - 1) S_i^2. \end{aligned}$$

where:

$SQT = (n-1)s_y^2$  is the **overall** variability of the  $n$  observations of  $Y$ ;

$SQF$  measures the **variability between** different factor levels;

$SQRE$  measures the **variability within** factor levels - and which cannot therefore be explained by the factor.

This is the **historical origin** of the name “the Analysis of Variance”: the variance of  $Y$  is broken up (“analysed”) into terms that are **associated with different causes**. In a one-way model, the causes are either the **factor** effects, or other (**residual**) causes that the one-factor model cannot explain.



# The summary table for one-way ANOVA

The information can be collected in an ANOVA summary table.

Source	d.f.	SQ	QM	$f_{calc}$
Factor	$k - 1$	$SQF = \sum_{i=1}^k n_i \cdot (\bar{y}_{i.} - \bar{y}_{..})^2$	$QMF = \frac{SQF}{k-1}$	$\frac{QMF}{QMRE}$
Residuals	$n - k$	$SQRE = \sum_{i=1}^k (n_i - 1) s_i^2$	$QMRE = \frac{SQRE}{n-k}$	
Total	$n - 1$	$SQT = (n - 1) s_y^2$	—	—

# One-way ANOVAs in

To carry out a one-way ANOVA in , we must organize the data in a `data.frame` with two columns:

- 1 one with the (numerical) values of the **response variable**;
- 2 another with the **factor** (specifying the factor level of each observation).

The formula used in `R` to specify the one-way ANOVA is similar to that used in a linear regression, indicating the factor name as the predictor.

For example, to carry out an ANOVA of petal widths over species, with the  $n = 150$  iris dataset, the formula is:

$$\text{Petal.Width} \sim \text{Species}$$

since the `iris` *data frame* has a column called `Species` which was defined as a factor.

## One-way ANOVAs in (cont.)

Although it is possible to use the command `lm` to request an ANOVA (ANOVAs being specific cases of the Linear Model), another command organizes the information in the more traditional way for ANOVAs: the command `aov`.

### One-way ANOVA (iris data, slide 230)

```
> aov(Petal.Width ~ Species, data=iris)
```

```
Call: aov(formula = Petal.Width ~ Species, data = iris)
```

```
Terms:
```

	Species	Residuals
Sum of Squares	80.41333	6.15660
Deg. of Freedom	2	147

```
Residual standard error: 0.20465
```

The output is different from that obtained with the well-known command `lm`.

## One-way ANOVAs in (cont.)

The command `summary`, when applied to a fitted ANOVA, produces the complete ANOVA summary table.

### One-way ANOVA (iris, slide 230)

```
> iris.aov <- aov(Petal.Width ~ Species , data=iris)
```

```
> summary(iris.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	80.413	40.207	960.01	< 2.2e-16 ***
Residuals	147	6.157	0.042		

In this case, the  $F$  test clearly rejects the hypothesis that the additive level effects,  $\alpha_i$ , are all zero. Thus, we reject the hypothesis that the mean petal widths are the same for all species.

Conclusion: the factor (species) affects the response variable (petal width).

## The estimated parameters, in

To extract the parameter estimates  $\mu_1, \alpha_2, \alpha_3, \dots, \alpha_k$ , the command `coef` can be applied to a fitted ANOVA model.

### One-way ANOVA (iris, slide 230)

```
> coef(iris.aov)
(Intercept) Speciesversicolor  Speciesvirginica
          0.246                1.080                1.780
```

These are the [estimated parameter values](#):

- $\hat{\mu}_1 = 0.246$ : sample mean of the *setosa* petal widths;
- $\hat{\alpha}_2 = 1.080$ : additive term which, when added to the *setosa* sample mean, gives the *versicolor* petal width sample mean;
- $\hat{\alpha}_3 = 1.780$ : additive term which, if added to the *setosa* sample mean, gives the *virginica* petal width sample mean.

## Estimated parameters in (cont.)

The **level means** of the response variable, can be obtained with the command `model.tables` and the argument `type="means"`:

### One-way ANOVA (iris, slide 230)

```
> model.tables(iris.aov , type="means")
```

```
Tables of means
```

```
Grand mean
```

```
1.199333
```

```
Species
```

```
Species
```

```
setosa versicolor virginica
```

```
0.246
```

```
1.326
```

```
2.026
```

 orders the factor level by alphabetical order.

# ANOVA as a Linear Model in

It is also possible to use the command `lm`, which is useful for inference on the model parameters:

## One-way ANOVA (iris, slide 230)

```
> summary(lm(Petal.Width ~ Species , data=iris))
Call: lm(formula = Petal.Width ~ Species, data = iris)
(...)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.24600    0.02894     8.50 1.96e-14 ***
Speciesversicolor  1.08000    0.04093    26.39 < 2e-16 ***
Speciesvirginica  1.78000    0.04093    43.49 < 2e-16 ***
--
Residual standard error: 0.2047 on 147 degrees of freedom
Multiple R-squared:  0.9289, Adjusted R-squared:  0.9279
F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

# Further exploration $H_1$

## Rejecting the Null Hypothesis

$$\alpha_2 = \alpha_3 = \dots = \alpha_k = 0 \Leftrightarrow \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

leaves open the issue of precisely which pairs of level means should be considered significantly different.

To determine for which pairs of levels  $i, j$  we should conclude that  $\mu_i \neq \mu_j$ , other tests are needed. **Multiple comparison tests** are advisable, so as to control the **overall significance level** of all  $\binom{k}{2}$  comparisons of pairs of means.

Among these, we highlight **Tukey's test** and **Scheffé's test**.

These are not covered in this course.



## Model checking in a one-way ANOVA

The validity of the model assumptions can be checked using similar approaches to those discussed in Linear Regression, with residual plots and other diagnostics to identify observations of particular impact. But there are some **specificities**.

In a one-way ANOVA plots of  $e_{ij}$  vs.  $\hat{y}_{ij}$ ,  $k$  columns of residuals appear, because the fitted values  $\hat{y}_{ij} = \bar{y}_i$  are the same for all observations from a given factor level.

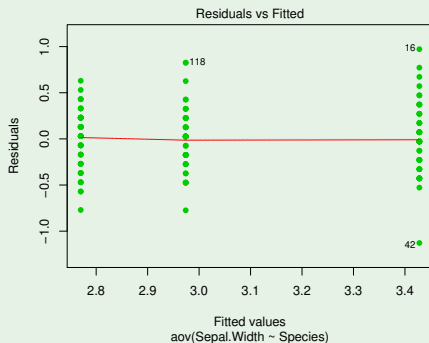
This pattern is **not** a violation of the model assumptions.

All observations from a common factor level will have the same leverage, equal to  $h_{ij} = \frac{1}{n_i}$ . Especially for balanced designs, leverages will be of little use in the context of ANOVAs.

# Residual plots in one-way ANOVAs (cont.)

## A residual plot in one-way ANOVA

```
> plot(aov(Sepal.Width ~ Species, data=iris), which=1)
```



# Violation of model assumptions in ANOVA

Violations of the model assumptions are not all equally serious. A few general comments:

- The ANOVA  $F$  test<sup>3</sup> is relatively robust to deviations from Normality.
- Violations of the assumption of variance homogeneity are, in general, less severe for balanced designs, but may be serious for strongly unbalanced designs.
- The lack of independence between random errors is the most serious violation of model assumptions and should be avoided, which is often possible with a suitable experimental design.

---

<sup>3</sup>And Tukey's multiple comparisons.

## A warning

In the **classical formulation** of the one-way ANOVA model, with equation

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \forall i, j$$

instead of imposing the condition  $\alpha_1 = 0$ , the alternative  $\sum_{i=1}^k \alpha_i = 0$  is used.

This alternative restriction:

- Changes the **interpretation** of the parameters: ( $\mu$  is now an **overall mean of  $Y$**  and  $\alpha_i$  the deviation of level mean  $\mu_i$  in relation to  $\mu$ );
- The parameter estimators change.
- The result of the  $F$  test for factor effects does **not** change.

Our choice of restriction,  $\alpha_1 = 0$ , besides being **extensive to models with more factors**, enables us to make direct use of the results studied for multiple linear regressions.

# Designs and Experimental units

When **designing experiments** to be analysed with ANOVAs or linear regressions, the observations of the response variable correspond to  $n$  different **experimental units** (individuals, plots of land, sites, etc.).

**General principles** in the selection of these experimental units are:

## Randomization

**Randomization**, that is, the **random selection** of experimental units and their association to a given factor level, when controllable. This is important to:

- **work with Probability Theory**; and
- **avoid bias** (even unwillingly).

## Repetitions

The **repetition** of **independent** observations is needed to **estimate the variability** associated with the estimation (standard errors) and **minimise the impact of outliers**.

# Repetitions and pseudo-repetitions

## Repetitions and pseudo-repetitions

A distinction must be made between **repetitions** and **pseudo-repetitions**. For example, in a study of tomato plants, there is a difference between:

- selecting two fruits from **the same plant**; or
- selecting two fruits from **different plants**.

The genotypes, phenotypes and environmental conditions of fruits from the **same plant** are identical or very similar. These are **pseudo-repetitions** (with correlated measurements), **not independent repetitions**.

Pseudo-repetitions **may be useful**: replacing each group of pseudo-repetitions by **a single mean observation** may **decrease the variability among different (independent) observations**, making inference more precise.

# Heterogeneity of experimental units

Variability in measurements on experimental units that is **not attributable to the predictors** is considered random variation and contemplated in the **random errors**. Thus, **uncontrolled heterogeneity** of the experimental units will increase the value of ***SQRE*** and ***QMRE***.

Increasing ***QMRE*** means that in a test for factor effects, **the computed value of the *F* statistic decreases**, drawing it away from the critical region. Hence,

## in an ANOVA

**uncontrolled heterogeneity** of experimental units contributes to **hide the presence of possible factor effects**.

## in a Linear Regression

**uncontrolled heterogeneity** of experimental units contributes to **worsen the quality of the model fit**, decreasing its  **$R^2$** .

# Controlling heterogeneity

Except for laboratory conditions, it is not possible to make experimental units fully homogeneous: the natural variability of plants, animals, soils, geographical conditions, cells, etc. means that uncontrolled variability of experimental units always exists.

Even if it were possible to have (nearly) homogeneous experimental units, there would be an undesirable drawback: results would only be valid for the type of experimental units used in the experiment.

When an important factor of variability of the experimental units is known to exist, the best way of controlling its effects is to **contemplate the existence of that factor of variability in the design and model**, so as to filter out its effects.



## An example

We seek to analyse the yield of 5 different varieties of wheat. Yields are also affected by soil types.

It is not always possible to have homogeneous soils in an experiment. Even if it were possible, it may not be desirable, because the validity of results would be restricted to that single type of soil.

Assume that we have four fields with different soil types. Each field may be split up into five plots of size viable for wheat.

Instead of associating the 5 varieties with the 20 plots totally at random, it is preferable to force each field to have one plot with each variety. Randomization will only be used within each field.

## An example (cont.)

The situation described on the previous slide:

Terreno 1 

Var.1	Var.3	Var.4	Var.5	Var.2
-------	-------	-------	-------	-------

Terreno 2 

Var.4	Var.3	Var.5	Var.1	Var.2
-------	-------	-------	-------	-------

Terreno 3 

Var.2	Var.4	Var.1	Var.3	Var.5
-------	-------	-------	-------	-------

Terreno 4 

Var.5	Var.2	Var.4	Var.1	Var.3
-------	-------	-------	-------	-------

There has been a **restriction to total randomization**: within each field there is randomization in the association of varieties to plots, but each field is forced to have one plot with each **variety**.

## Two-way (two-factor) factorial designs

The design discussed above is a specific case of a **two-way factorial design**, where one factor is **wheat variety** and a second factor is **soil type (field)**.

A **factorial design** is an experimental design in which **observations** are made for all possible combinations of levels from each factor.

Thus, designs with more than one factor may result from:

- the intention of actually studying possible effects of more than one factor on the response variable;
- an attempt to control experimental variability.

Historically, this second situation has given rise to the name **blocks**, and in the first situation we just speak of **factors**. But they are **analogous situations**.

## Two-way ANOVA model (without interaction)

We consider **two different ANOVA models** for a 2-way factorial design.

Consider:

- A response variable  $Y$ , on which  $n$  observations are collected.
- A Factor  $A$ , with  $a$  levels.
- A Factor  $B$ , with  $b$  levels.

A **first model** assumes the existence of **two different kinds of effects** on the values of  $Y$ : the effects associated with the levels of each factor.

# Representation of a two-way factorial design

		Factor B				
		$B_1$	$B_2$	$B_3$	...	$B_b$
Factor A	Levels $A_1$	× × ×	× × ×	× × ×	...	× × ×
	$A_2$	× × ×	× × ×	× × ×	...	× × ×
	$A_3$	× × ×	× × ×	× × ×	...	× × ×
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$A_a$	× × ×	× × ×	× × ×	...	× × ×

**Warning:** This representation does **not** correspond to any **spatial** organization of the experiment.

**Cell:** combines a level of one factor with a level of another factor. It corresponds to a given **experimental situation**.

In this design, there are  **$ab$**  experimental situations (cells), each with  **$n_{ij}$**  observations.

## Two-way ANOVA model (without interaction)

**Notation:** Each observation of the response variable is now identified by **three indices**,  $Y_{ijk}$ , where:

- $i$  indicates the **Factor A level  $i$**  ( $i = 1, 2, \dots, a$ ).
- $j$  indicates the **Factor B level  $j$**  ( $j = 1, 2, \dots, b$ ).
- $k$  indicates the  **$k$ -the repetition in cell  $(i, j)$**  ( $k = 1, 2, \dots, n_{ij}$ ).

The number of observations in cell  $(i, j)$  is represented by  $n_{ij}$ . We have:

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

If the number of observations is the same in every cell ( $n_{ij} = n_c, \quad \forall i, j$ ), we speak of a **balanced design**.

# A model equation for $Y$

A first model assumes that the expected value for each observation is given by:

$$E[Y_{ijk}] = \mu_{ij} = \mu + \alpha_i + \beta_j, \quad \forall i, j, k.$$

The parameter  $\mu$  is common to all observations.

Each parameter  $\alpha_i$  represents the increase associated with different levels of Factor A, and is called the effect of factor A level  $i$ .

Each parameter  $\beta_j$  represents the increase associated with different levels of Factor B, and is called the effect of factor B level  $j$ .

The variability of  $Y_{ijk}$  around its mean value is given by an additive random error,  $\varepsilon_{ijk}$ , with mean zero:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}.$$

# The model equation in vector notation

The model equation in a two-way ANOVA (without interaction effects) can also be written using vector notation.

Denote:

- $\vec{Y}$  the **random**  $n$ -dimensional vector with all observations of the response variable.
- $\vec{1}_n$  the vector of  $n$  ones.
- $\vec{I}_{A_i}$  the **indicator variable** for level  $i$  of Factor A.
- $\vec{I}_{B_j}$  the **indicator variable** for level  $j$  of Factor B.
- $\vec{\epsilon}$  the **random** vector with the  $n$  random errors.



# A first equation in vector notation

If we assume effects for **all** levels of both factors, the model equation will be:

$$\vec{Y} = \mu \vec{1}_n + \alpha_1 \vec{J}_{A_1} + \alpha_2 \vec{J}_{A_2} + \dots + \alpha_a \vec{J}_{A_a} + \beta_1 \vec{J}_{B_1} + \beta_2 \vec{J}_{B_2} + \dots + \beta_b \vec{J}_{B_b} + \vec{\epsilon}$$

The model matrix **X** defined by this model would have linearly dependent columns **for two reasons**:

- the sum of Factor A indicator variables is the column of ones,  $\vec{1}_n$ ;
- the sum of Factor B indicator variables is the column of ones,  $\vec{1}_n$ .

## A first model matrix $\mathbf{X}$

$$\mathbf{X} = \left[ \begin{array}{c|ccc|ccc|ccc}
 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\
 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\
 1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\
 1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\
 \hline
 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\
 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\
 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 1 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 1 \\
 1 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 1 \\
 \hline
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 \hline
 1 & 0 & 0 & \dots & 1 & 1 & 0 & \dots & 0 \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \\
 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1
 \end{array} \right]$$

$\uparrow \mathbf{1}_n$      $\uparrow \mathcal{J}_{A_1}$      $\uparrow \mathcal{J}_{A_2}$     ...     $\uparrow \mathcal{J}_{A_a}$      $\uparrow \mathcal{J}_{B_1}$      $\uparrow \mathcal{J}_{B_2}$     ...     $\uparrow \mathcal{J}_{B_b}$

Dropping the column  $\mathbf{1}_n$  does not overcome linear dependence.

## Restrictions on the model equation

Henceforth, we assume that the terms associated with the first level from each factor are dropped from the equation, that is:

$$\alpha_1 = 0 \quad \text{e} \quad \beta_1 = 0 ,$$

This corresponds to excluding columns  $\vec{\mathcal{J}}_{A_1}$  and  $\vec{\mathcal{J}}_{B_1}$  from matrix  $\mathbf{X}$ .

The two-way model equation in an ANOVA without interaction effects, is:

$$\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}}_n + \alpha_2 \vec{\mathcal{J}}_{A_2} + \dots + \alpha_a \vec{\mathcal{J}}_{A_a} + \beta_2 \vec{\mathcal{J}}_{B_2} + \dots + \beta_b \vec{\mathcal{J}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

The parameter  $\mu$  is now the expected value of  $Y$  for observations from cell  $(i=1, j=1)$ , and will be denoted as  $\mu_{11}$ :

$$Y_{11k} = \mu + \varepsilon_{11k} \quad \Rightarrow \quad E[Y_{11k}] = \mu = \mu_{11} .$$

# Model matrix in two-way ANOVA (no interaction)

$$\mathbf{X} = \begin{bmatrix}
 1 & 0 & \dots & 0 & 0 & \dots & 0 \\
 1 & 0 & \dots & 0 & 0 & \dots & 0 \\
 1 & 0 & \dots & 0 & 1 & \dots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 1 & 0 & \dots & 0 & 0 & \dots & 1 \\
 1 & 0 & \dots & 0 & 0 & \dots & 1 \\
 \hline
 1 & 1 & \dots & 0 & 0 & \dots & 0 \\
 1 & 1 & \dots & 0 & 0 & \dots & 0 \\
 1 & 1 & \dots & 0 & 0 & \dots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 1 & 1 & \dots & 0 & 0 & \dots & 1 \\
 1 & 1 & \dots & 0 & 0 & \dots & 1 \\
 \hline
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 \hline
 1 & 0 & \dots & 1 & 0 & \dots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 1 & 0 & \dots & 1 & 0 & \dots & 1 \\
 1 & 0 & \dots & 1 & 0 & \dots & 1 \\
 \hline
 \uparrow & \uparrow & & \uparrow & \uparrow & & \uparrow \\
 \vec{1}_n & \vec{J}_{A_2} & \dots & \vec{J}_{A_a} & \vec{J}_{B_2} & \dots & \vec{J}_{B_b}
 \end{bmatrix}$$

# The two-way ANOVA model, without interaction

We make the usual assumptions necessary for inference,

## Two-way ANOVA model, without interaction effects

Consider  $n$  observations,  $Y_{ijk}$ , of which  $n_{ij}$  correspond to cell  $(i, j)$  ( $i = 1, \dots, a$ ;  $j = 1, \dots, b$ ). Assume:

- 1  $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}$ ,  $\forall i=1, \dots, a; j=1, \dots, b; k=1, \dots, n_{ij}$  ( $\alpha_1 = 0$ ;  $\beta_1 = 0$ ).
- 2  $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ ,  $\forall i, j, k$
- 3  $\{\varepsilon_{ijk}\}_{i,j,k}$  independent random variables.

The model has  $a + b - 1$  unknown parameters:

- parameter  $\mu_{11}$ ;
- the  $a-1$  increases  $\alpha_i$  ( $i > 1$ ); and
- the  $b-1$  increases  $\beta_j$  ( $j > 1$ ).

# Testing for effects

The Null Hypothesis of a goodness-of-fit test is that **all** effects, whether for factor A or factor B, are simultaneously zero. **This does not distinguish the effects of each factor.**

It is more useful to **separately test for the existence of effects for each factor**:

- Test I:  $H_0 : \alpha_i = 0, \quad \forall i = 2, \dots, a ;$
- Test II:  $H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b.$

## Testing for Factor B effects

The model (vector) equation in a two-way ANOVA, without interaction (slide 285) is:

$$\vec{Y} = \mu \vec{1}_n + \alpha_2 \vec{\mathcal{J}}_{A_2} + \dots + \alpha_a \vec{\mathcal{J}}_{A_a} + \beta_2 \vec{\mathcal{J}}_{B_2} + \dots + \beta_b \vec{\mathcal{J}}_{B_b} + \vec{\epsilon}$$

Being a Linear Model, we can test the hypotheses:

$$H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b \quad \text{vs.} \quad H_1 : \exists j \text{ for which } \beta_j \neq 0 .$$

We use a **partial F test** comparing the **full model** with  $a + b - 1$  parameters:

$$\text{(Modelo } M_{A+B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk} ,$$

and the **submodel**, with  $a$  parameters and model equation:

$$\text{(Modelo } M_A) \quad Y_{ijk} = \mu_{11} + \alpha_i + \epsilon_{ijk} .$$

The latter is a **one-way ANOVA model** (with factor A).

# The test for Factor B effects

We can:

- build the matrices  $\mathbf{X}$  for the model ( $M_{A+B}$ ) and submodel ( $M_A$ ).
- Calculate the projection matrices  $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$  for each model.
- Obtain the fitted vectors  $\vec{\hat{Y}} = \mathbf{H}\vec{Y}$  and residual vectors  $\vec{\hat{E}} = (\mathbf{I} - \mathbf{H})\vec{Y}$  for each model.
- Obtain the Residual Sums of Squares,  $SQRE_{A+B}$  and  $SQRE_A$ .
- Carry out the partial  $F$  test, with test statistic:

$$\text{(Factor B Effects)} \quad F = \frac{\overbrace{SQRE_A - SQRE_{A+B}}^{=SQB}}{b-1}}{\frac{SQRE_{A+B}}{n-(a+b-1)}} = \frac{QMB}{QMRE}$$

$$\text{defining } QMB = \frac{SQB}{b-1} = \frac{SQRE_A - SQRE_{A+B}}{b-1}$$



# F for factor B effects

Given the two-way ANOVA model, without interaction effects:

## F Test for factor B effects

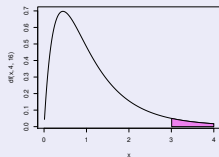
Hypotheses:  $H_0 : \beta_j = 0 \quad \forall j=2,\dots,b$  vs.  $H_1 : \exists j=2,\dots,b \text{ t.q. } \beta_j \neq 0$ .  
[B DOES NOT AFFECT Y] vs. [B AFFECTS Y]

Test statistic:  $F = \frac{QMB}{QMRE} \sim F_{(b-1, n-(a+b-1))}$  se  $H_0$ .

Significance level:  $\alpha$

Critical (Rejection) Region: One-sided, right tail

Reject  $H_0$  if  
 $F_{calc} > f_{\alpha(b-1, n-(a+b-1))}$



# Test statistic for Factor A effects

In a similar way, we have:

- $SQA = SQF_A$ , the Factor Sum of Squares in model  $M_A$ ;
- $QMA = \frac{SQA}{a-1}$ , the Factor Mean Square in model  $M_A$ ;
- $SQRE_{A+B}$  and  $QMRE = \frac{SQRE_{A+B}}{n-(a+b-1)}$  in model  $M_{A+B}$ .

The statistic

$$F = \frac{QMA}{QMRE} = \frac{\frac{SQA}{a-1}}{\frac{SQRE_{A+B}}{n-(a+b-1)}}$$

has an  $F_{[a-1, n-(a+b-1)]}$  distribution, when  $\alpha_i = 0$ , for all  $i=2, \dots, a$ .

# F test for factor A effects

Given the two-way ANOVA model, without interaction effects:

## F Test for factor A effects

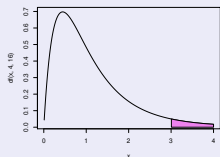
Hypotheses:  $H_0 : \alpha_j = 0 \quad \forall j=2, \dots, a$  vs.  $H_1 : \exists j=2, \dots, a \text{ t.q. } \alpha_j \neq 0$ .  
[A DOES NOT AFFECT Y] vs. [A AFFECTS Y]

Test Statistic:  $F = \frac{QMA}{QMRE} \sim F_{[a-1, n-(a+b-1)]}$  se  $H_0$ .

Significance level:  $\alpha$

Critical (Rejection) Region: One-sided, right-tail

Reject  $H_0$  if  
 $F_{calc} > f_{\alpha[a-1, n-(a+b-1)]}$



# The new decomposition of $SQT$

Considering the above-defined Sums of Squares, we have:

$$SQB = SQRE_A - SQRE_{A+B}$$

$$SQA = SQF_A = SQT - SQRE_A$$

Adding these SQs to  $SQRE_{A+B}$ , we get:

$$SQRE_{A+B} + SQA + SQB = SQT$$

which is a **new decomposition of  $SQT$** , into three terms, one for each factor effects and one for residual variability.

## Warning: Changing the order of factors

Exchanging the role of factors A and B defines Sums of Squares differently. Denoting by  $M_B$  the one-way ANOVA model, for factor B, we have:

$$\begin{aligned}SQB &= SQF_B = SQT - SQRE_B \\SQA &= SQRE_B - SQRE_{A+B}.\end{aligned}$$

It is still true that  $SQT$  can be decomposed as

$$SQT = SQA + SQB + SQRE_{A+B}.$$

Similar tests to those of slides 291 and 289 can be built.

But the two alternative definitions of  $SQA$  and  $SQB$  only give the same results for balanced designs. Only then will the order of the factors be arbitrary.

# SQA and SQB in balanced designs

In a **balanced design**, SQA and SQB are both Factor Sums of Squares of one-way ANOVA models (for factor A or B, slide 290).

Thus, in the formula for  $SQA = SQF_A$ , (slide 254), we have  $\hat{Y}_{ijk} = \bar{Y}_{i..}$  where  $\bar{Y}_{i..}$  indicates the mean of  $Y$  for factor A level  $i$ . Letting  $\bar{Y}_{...}$  be the overall mean of all  $n$  observations of  $Y$ , we have:

$$SQF_A = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (\hat{Y}_{ijk} - \bar{Y}_{...})^2 = bn_c \cdot \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2 = SQA.$$


In the same way,  $SQB = SQF_B$  is given by the fitted values of Model  $M_B$ , with Factor B only, where  $\hat{Y}_{ijk} = \bar{Y}_{.j.}$ . Thus:

$$SQF_B = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (\hat{Y}_{ijk} - \bar{Y}_{...})^2 = an_c \cdot \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2 = SQB.$$


# The 2-way ANOVA summary table (without interaction; balanced design)

Source	d.f.	SQ	QM	$f_{calc}$
Factor A	$a - 1$	$SQA = bn_c \cdot \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2$	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b - 1$	$SQB = an_c \cdot \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2$	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Residuals	$n - (a + b - 1)$	$SQRE = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (y_{ijk} - \hat{y}_{ijk})^2$	$QMRE = \frac{SQRE}{n - (a + b - 1)}$	
Total	$n - 1$	$SQT = (n - 1) s_y^2$	—	—

## Two-way ANOVA, without interaction in

To carry out a two-way ANOVA (without interaction) in , the data should be stored in a `data.frame` with three columns:

- 1 one for the (numerical) values of the **response variable**;
- 2 another for **factor A** (specifying the factor level for each observation);
- 3 the third for **factor B** (specifying its factor levels).

The formula used in  to specify a two-way ANOVA without interaction, is similar to that used in the two predictor Linear Regression, with the name of both factors separated by the symbol `+`:

$$y \sim fA + fB$$



# An example

## immer barley data (*package* MASS)

The yield of five varieties (*manchuria*, *svansota*, *velvet*, *trebi* and *peatland*) was registered in six locations<sup>a</sup>. In each location one plot was associated (at random) with each variety.

```
> summary(aov(Y1 ~ Var + Loc, data=immer))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Var	4	2756.6	689.2	4.2309	0.01214 *
Loc	5	17829.8	3566.0	21.8923	1.751e-07 ***
Residuals	20	3257.7	162.9		

Effects are slightly significant for varieties and strongly significant for locations. What about a one-way model ignoring locations?

```
> summary(aov(Y1 ~ Var, data=immer))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Var	4	2756.6	689.2	0.817	0.5264
Residuals	25	21087.6	843.5		

---

<sup>a</sup>Dados em Immer, Hayes & LeRoy Powers, Statistical adaptation of barley varietal adaptation, Journal of the American Society for Agronomy, 26, 403-419, 1934.

## Interpreting the parameter $\mu$

Interpreting the meaning of the model parameters depends on the convention used to overcome the issue of multicollinearity in the columns of matrix  $\mathbf{X}$ .

How can we interpret the parameters if we use the convention  $\alpha_1 = \beta_1 = 0$ ?

An observation of  $Y$  in cell  $(1, 1)$ , associated with crossing the first level of each factor, is of the form:

$$Y_{11k} = \mu_{11} + \underbrace{\alpha_1}_{=0} + \underbrace{\beta_1}_{=0} + \varepsilon_{11k} \quad \implies \quad E[Y_{11k}] = \mu_{11}$$

Parameter  $\mu_{11}$  corresponds to the expected value of the response variable  $Y$  in that cell (whose indicator variables were excluded from the model matrix).

## Interpreting the parameters $\alpha_j$

An observation of  $Y$  in cell  $(i, 1)$ , with  $i > 1$  (combining a factor A level different from the first, with the first level of Factor B) is of the form:

$$Y_{i1k} = \mu_{11} + \alpha_i + \underbrace{\beta_1}_{=0} + \varepsilon_{i1k} \quad \Longrightarrow \quad \mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$$

Parameter  $\alpha_i = \mu_{i1} - \mu_{11}$  is the increase in the expected value of the response variable  $Y$  associated with observations from level  $i > 1$  of Factor A (compared to  $\mu_{11}$ ), when  $j=1$ . It is called the **effect of factor A level  $i$** .

# Interpreting the parameters $\alpha_j$

Table with cell population means (means for each experimental situation):

		Factor B				
		$B_1$	$B_2$	$B_3$	...	$B_b$
Factor A	Levels $A_1$	$\mu_{11}$	$\mu_{12}$	$\mu_{13}$	...	$\mu_{1b}$
	$A_2$	$\mu_{21} = \mu_{11} + \alpha_2$	$\mu_{22}$	$\mu_{23}$	...	$\mu_{2b}$
	$A_3$	$\mu_{31} = \mu_{11} + \alpha_3$	$\mu_{32}$	$\mu_{33}$	...	$\mu_{3b}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$A_a$	$\mu_{a1} = \mu_{11} + \alpha_a$	$\mu_{a2}$	$\mu_{a3}$	...	$\mu_{ab}$

## Interpreting the parameters $\beta_j$

An observation of  $Y$  from cell  $(1, j)$ , with  $j > 1$  (combining the first Factor A level with a level of Factor B different from the first) is given by:

$$Y_{1jk} = \mu_{11} + \underbrace{\alpha_1}_{=0} + \beta_j + \varepsilon_{1jk} \quad \implies \quad \mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$$

The parameter  $\beta_j = \mu_{1j} - \mu_{11}$  is the increase in the expected value of the response variable  $Y$ , for observations from Factor B level  $j$  (compared to  $\mu_{11}$ ), when  $i=1$ . It is called the **effect of factor B level  $j$** .

# Interpreting the parameters $\beta_j$

Table with the cell population means (means for each experimental situation):

		Factor B				
		$B_1$	$B_2$	$B_3$	...	$B_b$
Factor A	Levels $A_1$	$\mu_{11}$	$\mu_{12} = \mu_{11} + \beta_2$	$\mu_{13} = \mu_{11} + \beta_3$	...	$\mu_{1b} = \mu_{11} + \beta_b$
	$A_2$	$\mu_{21}$	$\mu_{22}$	$\mu_{23}$	...	$\mu_{2b}$
	$A_3$	$\mu_{31}$	$\mu_{32}$	$\mu_{33}$	...	$\mu_{3b}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$A_a$	$\mu_{a1}$	$\mu_{a2}$	$\mu_{a3}$	...	$\mu_{ab}$

## Observations of $Y$ in a general situation

But this model is too rigid: there are no parameters left, hence the expected values in the remaining cells are already pre-determined.

For observations of  $Y$  in a generic cell  $(i, j)$ , with  $i > 1$  and  $j > 1$ , we have:

$$Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \quad \implies \quad \mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j.$$

All the terms in these expressions for the expected values of  $Y$  have already been used. There is no flexibility to describe the specific situation in cells with  $i > 1$  and  $j > 1$ .

A model without interaction effects is used above all when there is a single observation in each cell, i.e.,  $n_{ij} = 1, \forall i, j$ .

# Formulas for balanced designs

Denote:

$\bar{Y}_{i..}$  sample mean of the  $bn_c$  observations for level  $i$  of factor A,

$$\bar{Y}_{i..} = \frac{1}{bn_c} \sum_{j=1}^b \sum_{k=1}^{n_c} Y_{ijk}$$

$\bar{Y}_{.j.}$  sample mean of the  $an_c$  observations for level  $j$  of Factor B,

$$\bar{Y}_{.j.} = \frac{1}{an_c} \sum_{i=1}^a \sum_{k=1}^{n_c} Y_{ijk}$$

$\bar{Y}_{...}$  overall sample mean of all  $n = abn_c$  observations,

$$\bar{Y}_{...} = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} Y_{ijk}.$$

If the design is balanced, that is,  $n_{ij} = n_c, \forall i, j$ , we have:

- $\hat{\mu}_{11} = \bar{Y}_{1..} + \bar{Y}_{.1.} - \bar{Y}_{...}$
- $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{1..}$
- $\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{.1.}$



## Formulas for balanced designs (cont.)

Taking into account these formulas and the Model equation, the fitted values for each observation depend only on the overall mean and on the means of the observations in the corresponding level means for each factor:

$$\hat{Y}_{ijk} = \hat{\mu}_{11} + \hat{\alpha}_i + \hat{\beta}_j = \bar{Y}_{i..} + \bar{Y}_{.j.} - \bar{Y}_{...}, \quad \forall i, j, k$$

**Warning:** Unlike what happens in a one-way ANOVA, the fitted values  $\hat{Y}_{ijk}$  are **not** the mean of the observations of  $Y$  in the same experimental situation (cell  $(i, j)$ ).

## Models with interaction effects

When there are repetitions in the cells, the most natural way to model a two-factor design is to envisage the existence of a third kind of effects: **interaction effects**.

The idea is to include in the model equation for  $Y_{ijk}$  a term  $(\alpha\beta)_{ij}$  allowing each cell to have a specific effect associated with the combination of levels  $i$  of Factor A and  $j$  of Factor B:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} .$$

The effects  $\alpha_i$  and  $\beta_j$  are now called **main** effects for each Factor.

# The expected values of $Y_{ijk}$ (model with interaction)

We impose the following restrictions on the parameters:

$$\alpha_1 = 0 \quad ; \quad \beta_1 = 0 \quad ; \quad (\alpha\beta)_{1j} = 0, \forall j \quad ; \quad (\alpha\beta)_{i1} = 0, \forall i.$$

Levels		Factor B				
		$B_1$	$B_2$	$B_3$	...	$B_b$
Factor A	$A_1$	× × ×	× × ×	× × ×	...	× × ×
	$A_2$	× × ×	× × ×	× × ×	...	× × ×
	$A_3$	× × ×	× × ×	× × ×	...	× × ×
	⋮	⋮	⋮	⋮	⋮	⋮
	$A_a$	× × ×	× × ×	× × ×	...	× × ×

Only observations **not** from row 1 and/or column 1 have **interaction effect terms**.

Only observations **not** associated with  $A_1$  have **main effects**  $\alpha_j$ .

Only observations **not** associated with  $B_1$  have **main effects**  $\beta_j$ .

# The expected values of $Y_{ijk}$ (model with interaction)

With the restrictions, we have:

- For the first cell ( $i = j = 1$ ):  $\mu_{11} = E[Y_{11k}] = \mu$ .
- In the remaining cells  $(1, j)$  of the first level of Factor A:  
 $\mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$ .
- In the remaining cells  $(i, 1)$  of the first level of Factor B:  
 $\mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$ .
- In the remaining generic cells  $(i, j)$ , with  $i > 1$  and  $j > 1$ ,  
 $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ .

This model has  $ab$  parameters:

- one mean of the reference cell,  $\mu_{11}$ ;
- $a-1$  main effects for factor A,  $\alpha_i$  ( $i > 1$ );
- $b-1$  main effects for factor B,  $\beta_j$  ( $j > 1$ );
- $(a-1)(b-1)$  interaction effects,  $(\alpha\beta)_{ij}$  ( $i > 1, j > 1$ ).

# Cell indicator variables

The equation of the two-way ANOVA model, with interaction, is defined using **cell indicator variables** when  $i > 1$  and  $j > 1$ ,  $\vec{\mathcal{I}}_{A_i:B_j}$ :

$$\begin{aligned}\bar{\mathbf{Y}} = & \mu \bar{\mathbf{1}}_n + \alpha_2 \vec{\mathcal{I}}_{A_2} + \dots + \alpha_a \vec{\mathcal{I}}_{A_a} + \beta_2 \vec{\mathcal{I}}_{B_2} + \dots + \beta_b \vec{\mathcal{I}}_{B_b} + \\ & + (\alpha\beta)_{22} \vec{\mathcal{I}}_{A_2:B_2} + (\alpha\beta)_{23} \vec{\mathcal{I}}_{A_2:B_3} + \dots + (\alpha\beta)_{ab} \vec{\mathcal{I}}_{A_a:B_b} + \bar{\boldsymbol{\epsilon}}\end{aligned}$$

The model matrix  $\mathbf{X}$  now has  $ab$  columns:

- one column of ones,  $\bar{\mathbf{1}}_n$ , associated with parameter  $\mu_{11}$ .
- $a-1$  columns with level indicators for factor A,  $\vec{\mathcal{I}}_{A_i}$ , ( $i > 1$ ), associated with parameters  $\alpha_i$ .
- $b-1$  columns with level indicators for factor B,  $\vec{\mathcal{I}}_{B_j}$ , ( $j > 1$ ), associated with parameters  $\beta_j$ .
- $(a-1)(b-1)$  columns with cell indicators,  $\vec{\mathcal{I}}_{A_i:B_j}$ , ( $i, j > 1$ ), associated with interaction effects  $(\alpha\beta)_{ij}$ .

# The three ANOVA tests

In this design, we wish to test the existence of each of three kinds of effects:

- $H_0 : (\alpha\beta)_{ij} = 0, \quad \forall i = 2, \dots, a, \forall j = 2, \dots, b ;$
- $H_0 : \alpha_i = 0, \quad \forall i = 2, \dots, a ;$  e
- $H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b .$

The test statistics for each of these tests result from decomposing the Total Sum of Squares into suitable terms.

As in previous models,  $\vec{\hat{Y}} = \mathbf{H}\vec{Y}$ , with  $\mathbf{H}$  the 'hat' matrix that orthogonally projects onto the space  $\mathcal{C}(\mathbf{X})$  spanned by the columns of matrix  $\mathbf{X}$ .

And also:  $SQRE = \|\vec{Y} - \vec{\hat{Y}}\|^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2.$

# The two-way ANOVA model, with interaction

Adding the assumptions necessary for inference,

## Two-way ANOVA model, with interaction (Model $M_{A*B}$ )

Assume  $n$  observations,  $Y_{ijk}$ ,  $n_{ij}$  of which are associated with cell  $(i, j)$  ( $i = 1, \dots, a; j = 1, \dots, b$ ). We have:

- 1  $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ ,  $\forall i=1, \dots, a; j=1, \dots, b; k=1, \dots, n_{ij}$   
with  $\alpha_1 = 0; \beta_1 = 0; (\alpha\beta)_{ij} = 0$  if  $i = 1$  and/or  $j = 1$ .
- 2  $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$
- 3  $\{\varepsilon_{ijk}\}_{i,j,k}$  independent random variables.

The model has  $ab$  parameters.

# Testing interaction effects

To test the existence of interaction effects,

$$H_0 : (\alpha\beta)_{ij} = 0, \quad \forall i = 2, \dots, a, \quad \forall j = 2, \dots, b,$$

we carry out a **partial  $F$  test** comparing the model

$$\text{(Model } M_{A*B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

with the submodel (two-way, without interaction effects):

$$\text{(Model } M_{A+B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk},$$

The **Interaction Sum of Squares** is defined as the difference:

$$SQAB = SQRE_{A+B} - SQRE_{A*B}$$



# Testing the main effects for each Factor

To test for Factor B main effects,  $H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b$ , consider the models

$$\begin{array}{ll} \text{(Model } M_{A+B}) & Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \\ \text{(Model } M_A) & Y_{ijk} = \mu_{11} + \alpha_i + \varepsilon_{ijk}, \end{array}$$

and take:

$$SQB = SQRE_A - SQRE_{A+B}$$

$$SQA = SQF_A = SQT - SQRE_A$$

**Note:** These two Sums of Squares are defined just as in the model without interaction effects.

# The decomposition of $SQT$

We defined:

$$SQAB = SQRE_{A+B} - SQRE_{A*B}$$

$$SQB = SQRE_A - SQRE_{A+B}$$

$$SQA = SQF_A = SQT - SQRE_A$$

Adding these Sums of Squares to  $SQRE_{A*B}$ , we get:

$$SQRE_{A*B} + SQAB + SQA + SQB = SQT$$

This **decomposition of  $SQT$**  generates the quantities upon which the three test statistics associated with Model  $M_{A*B}$  are based.

# The summary table

Based on the decomposition on slide 314 we can build the **summary table for the two-way ANOVA, with interaction**.

Source	d.f.	SQ	QM	$f_{calc}$
Factor A	$a - 1$	SQA	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b - 1$	SQB	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Interaction	$(a - 1)(b - 1)$	SQAB	$QMAB = \frac{SQAB}{(a-1)(b-1)}$	$\frac{QMAB}{QMRE}$
Residuals	$n - ab$	SQRE	$QMRE = \frac{SQRE}{n-ab}$	
Total	$n - 1$	$SQT = (n - 1) s_y^2$	-	-

The **degrees of freedom for each kind of effect** are the number of parameters of that kind that remain after the restrictions are imposed.

The **residuals degree of freedom** are the number of observations ( $n$ ) minus the number of model parameters ( $ab$ ).

# The $F$ test for interaction effects

Assuming the two-way ANOVA Model, with interaction:

## $F$ Test for interaction effects

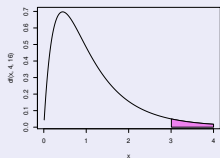
Hypotheses:  $H_0 : (\alpha\beta)_{ij} = 0 \quad \forall i, j$  vs.  $H_1 : \exists i, j \text{ s.t. } (\alpha\beta)_{ij} \neq 0$ .  
[NO INTERACTION] vs. [INTERACTION]

Test statistics:  $F = \frac{QMAB}{QMRE} \sim F_{[(a-1)(b-1), n-ab]}$  if  $H_0$ .

Significance Level:  $\alpha$

Critical (Rejection) Region: One-sided, right tail

Reject  $H_0$  if  
 $F_{calc} > f_{\alpha[(a-1)(b-1), n-ab]}$



# The $F$ Test for factor A main effects

Assuming the two-way ANOVA Model with interaction effects, we have:

## $F$ Test for factor A main effects

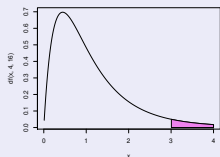
Hypotheses:  $H_0 : \alpha_j = 0 \quad \forall j=2, \dots, a$  vs.  $H_1 : \exists j=2, \dots, a \text{ t.q. } \alpha_j \neq 0$ .  
[NO FACTOR A EFFECTS] vs. [FACTOR A EFFECTS]

Test statistic:  $F = \frac{QMA}{QMRE} \sim F_{[a-1, n-ab]}$  if  $H_0$ .

Significance level:  $\alpha$

Critical (Rejection) Region: One-sided, right tail

Reject  $H_0$  if  
 $F_{calc} > f_{\alpha[a-1, n-ab]}$



# F Test for factor B main effects

Assuming the two-way ANOVA model with interaction effects, we have:

## F Test for factor B main effects

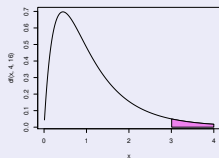
Hypotheses:  $H_0 : \beta_j = 0 \quad \forall j=2, \dots, b$  vs.  $H_1 : \exists j=2, \dots, b \text{ t.q. } \beta_j \neq 0.$   
[NO FACTOR B EFFECTS] vs. [FACTOR B EFFECTS]

Test statistic:  $F = \frac{QMB}{QMRE} \sim F_{[b-1, n-ab]}$  if  $H_0$ .


Significance level:  $\alpha$

Critical (Rejection) Region: One-sided, right tail

Reject  $H_0$  if  
 $F_{calc} > f_{\alpha[b-1, n-ab]}$



## Two-way ANOVAs, with interaction, in

For a two-way ANOVA with interaction in , organize the data as for the model without interaction: a **three-column data frame**,

- 1 one for the response variable;
- 2 one for factor A;
- 3 one for factor B.

The formula used in  to specify a two-way ANOVA with interaction, uses the symbol **\***:

$$y \sim fA * fB$$

where  $y$  is the response variable and  $fA$  and  $fB$  are the factor names.

# An example of a two-way model with interaction

## Data: yields of Negra Mole grape variety

A study to select genotypes of the Negra Mole grape variety (factor `clone`) with good yields (response variable `rend`) throughout the years (factor `ano`).

```
> NegraMole.aov <- aov(rend ~ ano*clone, data=NegraMole)
```

```
> summary(NegraMole.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ano	4	203.61	50.90	77.460	< 2e-16 ***
clone	6	26.39	4.40	6.694	1.41e-06 ***
ano:clone	24	18.08	0.75	1.146	0.294
Residuals	245	161.00	0.66		

There are clear `ano` and `clone` effects. Interaction effects are not significant. In the selection of genotypes, this is good (predictable behaviour).



# Again the Negra Mole example

## Yield data for Negra Mole variety

The overall, yearly, genotype and cell (year  $\times$  genotype combination) means can be obtained with the command `model.tables`.

```
> model.tables(NegraMole.aov, type="means")
```

```
Tables of means
```

```
Grand mean
```

```
2.2237          <-- overall mean yield
```

```
ano
```

```
LOU94 LOU95 LOU96 LOU97 LOU98
```

```
1.033 2.786 3.378 2.425 1.496      <-- 96 a good year, 94 and 98 bad
```

```
clone
```

```
NM0307 NM0507 NM0703 NM1006 NM2001 NM2015 NM2102
```

```
2.4410 1.7295 2.2294 1.8306 2.2362 2.5246 2.5747  <-- there are significant differences
```

```
ano:clone
```

```
clone
```

```
ano      NM0307 NM0507 NM0703 NM1006 NM2001 NM2015 NM2102
```

```
LOU94  1.465  0.710  0.675  0.814  1.409  0.949  1.209  <-- A bad year is bad for all genotypes
LOU95  2.994  2.290  2.783  2.310  2.557  3.619  2.947
LOU96  3.786  2.784  3.472  2.653  3.205  3.587  4.160  <-- A good year is good for all genotypes
LOU97  2.728  1.728  2.667  2.272  2.547  2.205  2.831  What happens in each cell is fairly
LOU98  1.233  1.135  1.550  1.105  1.463  2.263  1.727  predictable, since there is no interaction
```

## (No) Visualization of interaction effects

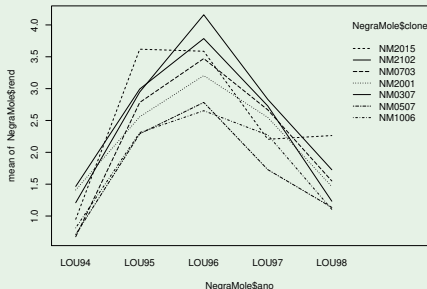
The existence of **interaction effects** may be seen in **plots** where:

- The **horizontal axis** is associated with levels of **one factor** (e.g.,  $fA$ );
- the **vertical axis** has mean values of the **response variable**  $Y$  in each cell;
- **for each cell**, a **point** is drawn, with coordinates given by the level of the factor and the respective cell mean of the response variable;
- **line segments** are used to unite the **points corresponding to the same level of the other factor** (e.g.,  $fB$ ).

# (No) Interaction plots in R

## Interaction plot for Negra Mole

```
> attach(NegraMole)
> interaction.plot(x.factor=ano, trace.factor=clone, response=rend)
> detach(NegraMole)
```



The absence of significant interaction translates into “approximately parallel curves”.

# Data from Exercise ANOVA 7 (sapotis)

**Response variable:** tanine content in the pulp

**Factor:** Storage temperature (high/low)

**Factor:** Storage time (0/3/6/9 days)

## Sapoti data (Exercise ANOVA 7)

```
> sapoti.aov <- aov(taninos ~ temperatura * tempo , data=sapoti)
> summary(sapoti.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
temperatura	1	206.0	206.0	238.6	5.72e-14	***
tempo	3	288.0	96.0	111.2	3.27e-14	***
temperatura:tempo	3	968.0	322.7	373.7	< 2e-16	***
Residuals	24	20.7	0.9			

All types of effects are clearly significant.

(No) With significant interaction, lines in an interaction plot will be far from parallel

# Exercise ANOVA 7 (cont.)

## Tanine content in sapotis

The overall, per storage temperature, per storage time and cell means can be obtained with the command `model.tables`.

```
> model.tables(sapoti.aov , type="means")
```

```
Tables of means
```

```
Grand mean
```

```
22.14375 <-- overall mean tanine content
```

```
temperatura
```

```
alta baixa
```

```
24.681 19.606 <-- mean tanine content for each storage temp
```

```
tempo
```

```
0 3 6 9
```

```
25.862 23.825 20.987 17.900 <-- mean tanine content for each storage time
```

```
temperatura:tempo
```

```
tempo
```

```
temperatura 0 3 6 9
```

```
alta 19.50 26.85 25.97 26.40
```

```
<-- usually larger, but not for zero storage
```

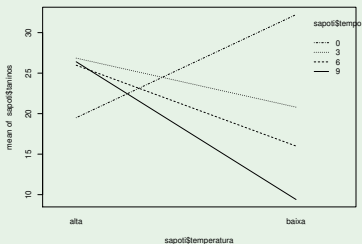
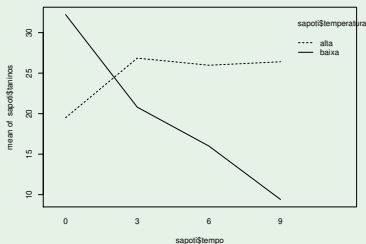
```
baixa 32.22 20.80 16.00 9.40
```

```
<-- usually smaller, but not for zero storage
```

# (No) Interaction plot

## Sapoti Data (Exercise ANOVA 7)

```
> attach(sapoti)
> interaction.plot(response=taninos,x.factor=tempo,trace.factor=temperatura)
> interaction.plot(response=taninos,x.factor=temperatura,trace.factor=tempo)
> detach(sapoti)
```



The significance of interaction effects must always be assessed with the corresponding  $F$  test.

# The study of interaction needs repetitions

In order to study interaction effects, it is necessary to have repetitions within cells.

The degrees of freedom of  $SQRE$  in this model are  $n - ab$ . With a single observation in each cell, we have  $n = ab$ , that is, as many parameters as observations. In this case, it is not even possible to define the Residual Mean Square,  $QMRE = \frac{SQRE}{n - ab}$ .

In a design without cell repetitions, only a model without interaction effects can be fitted. With repetitions, a model with interaction effects is the natural choice.

# Fitted values of $Y$ in a model with interaction effects

Let

$\bar{Y}_{ij}$  be the sample mean of the  $n_{ij}$  observations in cell  $(i,j)$ ,

$\bar{Y}_{i..}$  be the sample mean of the  $\sum_j n_{ij}$  observations of level  $i$  of Factor A,

$\bar{Y}_{.j}$  be the sample mean of the  $\sum_i n_{ij}$  observations of level  $j$  of Factor B,

$\bar{Y}_{...}$  be the overall sample mean of all  $n = \sum_i \sum_j n_{ij}$  observations.

The fitted values  $\hat{Y}_{ijk}$  are the same for all observations in a given cell, and are the cell sample mean:

$$\hat{Y}_{ijk} = \bar{Y}_{ij}.$$



## Parameter estimators

The estimators of the parameters in a two-way ANOVA model with interaction are:

- $\hat{\mu}_{11} = \bar{Y}_{11}.$
- $\hat{\alpha}_i = \bar{Y}_{i1.} - \bar{Y}_{11.} \quad (i > 1)$
- $\hat{\beta}_j = \bar{Y}_{1j.} - \bar{Y}_{11.} \quad (j > 1)$
- $(\hat{\alpha}\hat{\beta})_{ij} = (\bar{Y}_{ij.} + \bar{Y}_{11.}) - (\bar{Y}_{i1.} + \bar{Y}_{1j.}) \quad (i, j > 1).$

Confidence Intervals or Hypothesis Tests for individual parameters or linear combinations of the parameters can be carried out using the general theory of Linear Models.

# Residual Sum of Squares

Since the fitted values for each observation are their cell means,  $\hat{Y}_{ijk} = \bar{Y}_{ij}$ , we have:

$$SQRE = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij})^2$$

$$\Leftrightarrow SQRE = \sum_{i=1}^a \sum_{j=1}^b (n_{ij} - 1) S_{ij}^2,$$

where  $S_{ij}^2$  is the sample variance of the observations in cell  $(i, j)$ .

In a **balanced design**, we have  $n = n_c ab$ , and the **Residual Mean Square** is the simple mean of the cell sample variances,  $S_{ij}^2$ :

$$QMRE = \frac{SQRE}{n - ab} = \frac{n_c \cancel{ab} \uparrow}{ab \cancel{(n_c \downarrow)}} \sum_{i=1}^a \sum_{j=1}^b S_{ij}^2 = \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b S_{ij}^2.$$

## Other SQs for balanced designs

For **balanced designs** (with  $n_c$  observations per cell), it is also possible to obtain simple formulas for the **Sums of Squares associated with the main effects of each factor**.

These formulas are equal to those for the same Sums of Squares in a model without interaction effects:

$$SQA = bn_c \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$$

$$SQB = an_c \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

## A warning

In the classical formulation of the two-way ANOVA model with interaction effects, with model equation  $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ , instead of imposing the constraints  $\alpha_1 = \beta_1 = (\alpha\beta)_{11} = (\alpha\beta)_{1j} = 0$  ( $\forall i, j$ ), effects of all kinds, for any value of  $i$  and  $j$ , are assumed and the following alternative constraints are imposed:

- $\sum_i \alpha_i = 0$ ;
- $\sum_j \beta_j = 0$ ;
- $\sum_i (\alpha\beta)_{ij} = 0$ ,  $\forall j$ ;
- $\sum_j (\alpha\beta)_{ij} = 0$ ,  $\forall i$ .

These alternative constraints:

- change the interpretation of the parameters;
- change the parameter estimates;
- **do not** change the result of the  $F$  tests for the existence of effects.

## (No) Final comments on ANOVA

1. Um delineamento factorial pode ser definido com qualquer número de factores.

Num delineamento factorial a três factores (Factores A, B e C, com a, b e c níveis) há  $abc$  situações experimentais, todas com observações.

Cada observação indexa-se com quatro índices:  $Y_{ijkl}$  indica a observação  $l$  na célula  $(i, j, k)$ . Na equação de base para  $Y_{ijkl}$  há sete tipos de efeitos:

- três efeitos principais de cada factor,  $\alpha_i$ ,  $\beta_j$  e  $\gamma_k$ .
- três efeitos de interação dupla associados a cada combinação de níveis de dois Factores diferentes:  $(\alpha\beta)_{ij}$ ,  $(\alpha\gamma)_{ik}$  e  $(\beta\gamma)_{jk}$ .
- um efeito de tripla interação nas células onde se cruzam níveis dos três factores:  $(\alpha\beta\gamma)_{ijk}$

Para evitar um excesso de parâmetros, Consideram-se nulos os efeitos em que pelo menos um índice é igual a 1.

## (No) 1. O modelo factorial a três factores

A equação de base do modelo é agora da forma:

$$Y_{ijkl} = \mu_{1111} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} .$$

Com as restrições, o modelo tem *abc* parâmetros.

A Soma de Quadrados Total é agora decomposta em *oito parcelas*:

$$SQT = SQA + SQB + SQC + SQAB + SQAC + SQBC + SQABC + SQRE .$$

As sete *SQs* associadas a efeitos são *definidas pela diferença das Somas de Quadrados Residuais de modelos onde se vão sucessivamente omitindo os efeitos correspondentes*.

Há *sete testes*: um para cada tipo de efeitos. As estatísticas dos sete testes são todas do tipo  $F = \frac{QM_x}{QMRE}$ , onde *x* designa o tipo de efeitos a ser testado.

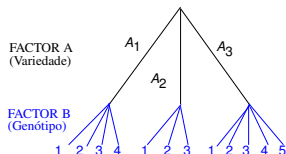
## (No) 2. Nested designs

São delineamentos com dois (ou mais) factores, em que os níveis de um dos factores variam consoante os níveis do outro factor.

Exemplo: dois factores, variedades e génotipos.

Um delineamento factorial é impossível.

Mas pode considerar-se uma estrutura hierárquica, representada no dendrograma à direita.



A equação base do modelo inclui efeitos de nível do Factor A e efeitos de nível do factor B, subordinado a A:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} .$$

Não faz sentido falar em efeitos do nível  $j$  do Factor B, sem especificar qual o nível do Factor A a que nos referimos. Nem faz sentido falar em efeitos de interacção: os níveis de cada factor não são, em geral, cruzados.

Haverá agora dois testes  $F$ : um para cada tipo de efeitos ( $\alpha_i$  e  $\beta_{j(i)}$ ). As estatísticas de teste obtêm-se de forma análoga, a partir da decomposição  $SQT = SQA + SQB(A) + SQRE$ .

## (No) 3. Outros tipos de delineamentos experimentais

Existem numerosos outros tipos de delineamentos mais complexos.

Alguns delineamentos visam reduzir o número de situações experimentais que é necessário estudar.

Exemplo: **quadrados latinos** ou **greco-romanos**.

Outros delineamentos visam ultrapassar dificuldades práticas na execução de uma experiência, como é o caso dos delineamentos em **parcelas divididas** (*split plots*).



## (No) 4. Métodos não paramétricos de tipo ANOVA

Uma forma alternativa de estudar problemas análogos aos objectivos de ANOVAs resulta da utilização de **métodos não paramétricos**:

- Não exigem pressupostos tão restritivos como os métodos clássicos, (e.g., a Normalidade ou homogeneidade de variâncias).
- Em contrapartida têm menor capacidade de rejeitar as hipóteses nulas caso elas sejam falsas (i.e., têm menor **potência**), quando os pressupostos adicionais dos métodos clássicos são válidos.
- Frequentemente, substituem os valores observados da variável resposta pelas **ordens (ranks)** dessas observações. As estatísticas de teste são então funções dessas ordens.

## (No) 4. Métodos não paramétricos de tipo ANOVA (cont.)

O teste de Kruskal-Wallis é uma alternativa não paramétrica à ANOVA a 1 Factor, em que:

- A hipótese nula é que nos vários níveis do factor as observações seguem a mesma distribuição.
- A hipótese alternativa é que a distribuição dos vários níveis difere apenas nas suas localizações (medianas).
- Cada observação é substituída pela sua ordem;
- A estatística de teste compara as ordens médias em cada nível do factor com a ordem média global, havendo uma distribuição exacta e uma distribuição assintótica para grandes amostras.

O teste de Kruskal-Wallis é equivalente a um teste ANOVA a um Factor sobre as ordens das observações.

# Analysis of Covariance: an introduction

The Linear Regressions and Analyses of Variance studied so far are particular cases of the **Linear Model**, which also encompasses the **Analyses of Covariance**.

In all three contexts, we model a **numerical** response variable  $Y$ .

What sets the three situations apart is the nature of the predictors.

- In a **Linear Regression**, the predictors are also **numerical** variables.
- In an **Analysis of Variance**, predictors are **factors**.
- In **Analyses of Covariance**, among the predictors we find **both numerical variables and factors**.

# Comparing regression lines for different factor levels

The Analysis of Covariance will be discussed in a frequent specific context of practical interest, associated with Linear Regressions.

We seek to compare regression lines relating a numerical variable  $Y$  and a numerical predictor  $x$ , in different contexts defined by the levels of a given factor.

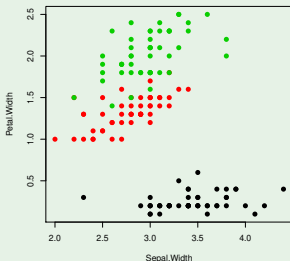
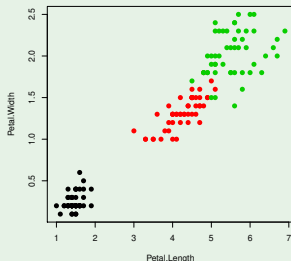
This, we have:

- a numerical response variable  $Y$ ;
- a numerical predictor  $x$ ;
- a factor predictor, that defines the different contexts for which we seek to compare the linear relation between  $Y$  and  $x$ .

# An example

## Predicting iris petal width considering species

Predicting petal width based on their length (left plot) produced a good model when the three species were pooled. And separately?



Predicting petal width based on sepal width (right plot) gives a bad model when the three species are pooled. And separately?

# ANCOVA as a comparison tool

We will formulate the problem assuming:

- an ANCOVA corresponding to a specific linear relation between  $Y$  and  $x$  for each factor level;
- different submodels correspond to assuming that some parameters are the same in those lines for different factor levels.

Being Linear Models, the available theory allows us to choose between the full model and any given submodel in this Analysis of Covariance.

We discuss the issue assuming (as in the example) that the factor has  $k = 3$  levels. But the approach can be also extended to any number  $k \in \mathbb{N}$  of levels.

# An Analysis of Covariance for the example

Assume a linear relation between response variable  $Y$  and predictor  $x$ , possibly different for each context defined by the **factor** (e.g., iris species):

- Context 1:  $Y = \beta_0 + \beta_1 x + \varepsilon$
- Context 2:  $Y = \beta_0^* + \beta_1^* x + \varepsilon$
- Context 3:  $Y = \beta_0^{**} + \beta_1^{**} x + \varepsilon$

Consider the first context as a **reference level** and write the parameters for other contexts using those of the reference level:

$$\begin{aligned} \beta_0^* &= \beta_0 + \alpha_{0:2} & ; & & \beta_1^* &= \beta_1 + \alpha_{1:2} \\ \beta_0^{**} &= \beta_0 + \alpha_{0:3} & ; & & \beta_1^{**} &= \beta_1 + \alpha_{1:3} \end{aligned}$$

With the parameters for each line written in this way, **the hypothesis that the three regression lines are the same is the hypothesis**

$$\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0 .$$

## The variables associated with the increases

Assume  $n$  observations, of which  $n_i$  from each level ( $i = 1, 2, 3$ ). As in a one-way ANOVA, use **double indexing** to identify levels of origin:  $Y_{ij}$  and  $x_{ij}$ .

Define **indicator variables**  $\vec{\mathcal{I}}_i$  for each level.

Define also **vectors with the values of the predictor  $x$  for a given level  $i$  ( $i > 1$ ) and zero in other positions**, which we represent as  $\vec{\mathbf{x}} \circ \vec{\mathcal{I}}_i$ .

In the earlier example with  $n_1 = 3$ ,  $n_2 = 4$  and  $n_3 = 2$  observations:

$$\vec{\mathcal{I}}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mathbf{x}} \circ \vec{\mathcal{I}}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{23} \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mathcal{I}}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad \vec{\mathbf{x}} \circ \vec{\mathcal{I}}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ x_{31} \\ x_{32} \end{bmatrix}$$



# The equation for the ANCOVA model

We can now write the relation between vector  $\vec{Y}$  of the  $n$  observations of the response variable and the predictor  $X$ , as follows:

$$\vec{Y} = \beta_0 \vec{1}_n + \beta_1 \vec{x} + \alpha_{0:2} \vec{\mathcal{J}}_2 + \alpha_{0:3} \vec{\mathcal{J}}_3 + \alpha_{1:2} (\vec{x} \circ \vec{\mathcal{J}}_2) + \alpha_{1:3} (\vec{x} \circ \vec{\mathcal{J}}_3) + \vec{\epsilon}.$$

In the example, using vector/matrix notation  $\vec{Y} = \mathbf{X}\beta + \vec{\epsilon}$ :

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & 0 & 0 & 0 & 0 \\ 1 & x_{12} & 0 & 0 & 0 & 0 \\ 1 & x_{13} & 0 & 0 & 0 & 0 \\ 1 & x_{21} & 1 & 0 & x_{21} & 0 \\ 1 & x_{22} & 1 & 0 & x_{22} & 0 \\ 1 & x_{23} & 1 & 0 & x_{23} & 0 \\ 1 & x_{24} & 1 & 0 & x_{24} & 0 \\ 1 & x_{31} & 0 & 1 & 0 & x_{31} \\ 1 & x_{32} & 0 & 1 & 0 & x_{32} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \alpha_{0:2} \\ \alpha_{0:3} \\ \alpha_{1:2} \\ \alpha_{1:3} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix}$$

# The ANCOVA model equation

The model on slide 345 fits a separate line for the observations in each context.

$$Y_{ij} = \begin{cases} \beta_0 + \beta_1 x_{1j} + \varepsilon_{1j}, & \text{se } i = 1 \\ (\beta_0 + \alpha_{0:2}) + (\beta_1 + \alpha_{1:2}) x_{2j} + \varepsilon_{2j}, & \text{se } i = 2 \\ (\beta_0 + \alpha_{0:3}) + (\beta_1 + \alpha_{1:3}) x_{3j} + \varepsilon_{3j}, & \text{se } i = 3. \end{cases} \quad (1)$$

If the parameters of type  $\alpha_{i:j}$  are *all* zero, the regression lines coincide, in all three contexts.

With the usual assumptions for random errors, this ANCOVA model is a **linear model** with  $3 \times 2 = 6$  parameters (and predictor variables  $\vec{x}$ ,  $\vec{\mathcal{I}}_2$ ,  $\vec{\mathcal{I}}_3$ ,  $\vec{x} \circ \vec{\mathcal{I}}_2$ ,  $\vec{x} \circ \vec{\mathcal{I}}_3$ ).

In general, for  $k$  factor levels there are  $2k$  parameters.

## Some interesting submodels

$$\vec{Y} = \beta_0 \vec{1}_n + \beta_1 \vec{x} + \alpha_{0:2} \vec{\mathcal{J}}_2 + \alpha_{0:3} \vec{\mathcal{J}}_3 + \alpha_{1:2} (\vec{x} \circ \vec{\mathcal{J}}_2) + \alpha_{1:3} (\vec{x} \circ \vec{\mathcal{J}}_3) + \vec{\epsilon}$$

- The hypothesis of a single regression line in the 3 contexts is  $\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0$ .
- The hypothesis of three parallel lines (i.e., equal slope), but possibly different intercepts, is  $\alpha_{1:2} = \alpha_{1:3} = 0$ .
- The hypothesis that the first and second lines have the same slope, is  $\alpha_{1:2} = 0$ .
- The hypothesis that the second and third lines have the same slope, is  $\alpha_{1:2} = \alpha_{1:3}$ , ou seja,  $\alpha_{1:2} - \alpha_{1:3} = 0$ .
- The hypothesis of three lines with the same intercept, but possibly different slopes, is  $\alpha_{0:2} = \alpha_{0:3} = 0$ .

These (or similar) hypotheses may be tested using the  $F$  and  $t$  tests discussed previously for linear models.

# Crossing factors and numerical variables in

In R, an ANCOVA model regressing  $y$  on  $x$ , allowing for different lines for each level of factor  $f$ , is specified by the formula:  $y \sim x * f$ .

## ANCOVA with the iris data

```
> modespecie.lm <- lm(Petal.Length ~ Sepal.Length * Species, data=iris)
> summary(modespecie.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.8031	0.5310	1.512	0.133
Sepal.Length	0.1316	0.1058	1.244	0.216
Speciesversicolor	-0.6179	0.6837	-0.904	0.368
Speciesvirginica	-0.1926	0.6578	-0.293	0.770
Sepal.Length:Speciesversicolor	0.5548	0.1281	4.330	2.78e-05 ***
Sepal.Length:Speciesvirginica	0.6184	0.1210	5.111	1.00e-06 ***

--

Residual standard error: 0.2611 on 144 degrees of freedom  
Multiple R-squared: 0.9789, Adjusted R-squared: 0.9781  
F-statistic: 1333 on 5 and 144 DF, p-value: < 2.2e-16

The regression line for *setosa* (reference level) has a significantly different slope from those for the other species.

## An example on . The 3 lines.

### ANCOVA with iris (cont.)

The three lines fitted by the ANCOVA model:

For species *setosa* (reference):

$$PL = 0.8031 + 0.1316 SL$$

For species *versicolor*:

$$PL = (0.8031 - 0.6179) + (0.1316 + 0.5548) SL = 0.1851 + 0.6865 SL$$

For species *virginica*:

$$PL = (0.8031 - 0.1926) + (0.1316 + 0.6184) SL = 0.6105 + 0.7501 SL$$

These are the same lines that result from a linear regression using only the  $n_j = 50$  observations from each species.

# The 3 separate linear regression lines

## ANCOVA with iris (cont.)

The three lines fitted by the ANCOVA model:

Species *Setosa*:  $PL = 0.8031 + 0.1316 SL$

Species *Versicolor*:  $PL = 0.1851 + 0.6865 SL$

Species *Virginica*:  $PL = 0.6105 + 0.7501 SL$

The three lines in separate linear regressions:

```
> coef(lm(Petal.Length ~ Sepal.Length , data=iris[1:50,]))
(Intercept) Sepal.Length
 0.8030518    0.1316317
> coef(lm(Petal.Length ~ Sepal.Length , data=iris[51:100,]))
(Intercept) Sepal.Length
 0.1851155    0.6864698
> coef(lm(Petal.Length ~ Sepal.Length , data=iris[101:150,]))
(Intercept) Sepal.Length
 0.6104680    0.7500808
```

# A block matrix $\mathbf{H}$ in ANCOVA

This is the result of the special structure of the matrix of orthogonal projections  $\mathbf{H}$ , associated with the ANCOVA model on slide 345.

Let  $\mathbf{H}_i$  be the  $n_i \times n_i$  orthogonal projection matrix onto subspace  $\mathcal{C}(X_{[i]}) \subset \mathbb{R}^{n_i}$  spanned only by the observations from level  $i$  of the factor.

The matrix  $\mathbf{H}$  in the ANCOVA model is then a **block-diagonal matrix** (assuming the rows of  $\mathbf{X}$  are grouped by factor level):

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{H}_k \end{bmatrix}$$

The fitted values of vector  $\vec{\mathbf{Y}} = \mathbf{H}\vec{\mathbf{Y}}$  depend only on the matrix  $\mathbf{H}_i$  for the corresponding level  $i$ .

## An example in . A single line?

### Is a single line for the three species admissible?

```
> modunico.lm <- lm(Petal.Length ~ Sepal.Length, data=iris)
```

```
> anova(modunico.lm, modespecie.lm)
```

Analysis of Variance Table

Model 1: Petal.Length ~ Sepal.Length

Model 2: Petal.Length ~ Sepal.Length \* Species

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	148	111.459				
2	144	9.818	4	101.641	372.7	< 2.2e-16 ***

We reject the hypothesis of a single line, in favour of different lines.



## Another example in $\mathbb{R}$ . Parallel lines?

In  $\mathbb{R}$ , a regression of  $y$  over  $x$  with parallel lines, but allowing for different intercepts for each level of factor  $f$ , is specified by the formula:  $y \sim x + f$

### Model for parallel lines, iris data

```
> modparalelas.lm <- lm(Petal.Length ~ Sepal.Length + Species, data=iris)
> summary(modparalelas.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-1.70234	0.23013	-7.397	1.01e-11	***
Sepal.Length	0.63211	0.04527	13.962	< 2e-16	***
Speciesversicolor	2.21014	0.07047	31.362	< 2e-16	***
Speciesvirginica	3.09000	0.09123	33.870	< 2e-16	***

--

Residual standard error: 0.2826 on 146 degrees of freedom  
Multiple R-squared: 0.9749, Adjusted R-squared: 0.9744  
F-statistic: 1890 on 3 and 146 DF, p-value: < 2.2e-16

## An example in : 3 parallel lines

### Parallel lines with iris data

The three lines fitted by the parallel lines model:

For species *setosa* (reference):

$$PL = -1.70234 + 0.63211 SL$$

For species *versicolor*:

$$PL = (-1.70234 + 2.21014) + 0.63211 SL = 0.50780 + 0.63211 SL$$

For species *virginica*:

$$PL = (-1.70234 + 3.09000) + 0.63211 SL = 1.38766 + 0.63211 SL$$

## An example in . Parallel lines? (cont.)

Is it admissible to assume that the three lines are parallel?

Let us perform a partial  $F$  test, comparing the submodel with parallel lines and the model assuming different lines.

### Partial $F$ test studying the parallel lines model

```
> anova(modparalelas.lm, modespecie.lm)
```

Analysis of Variance Table

Model 1: Petal.Length ~ Sepal.Length + Species

Model 2: Petal.Length ~ Sepal.Length \* Species

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	146	11.6571				
2	144	9.8179	2	1.8393	13.489	4.272e-06 ***

We reject the hypothesis of parallel lines.

# A warning about the assumptions

The previous tests are valid for the **assumptions of Linear Models**:

- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \forall i, j;$
- independent random errors.

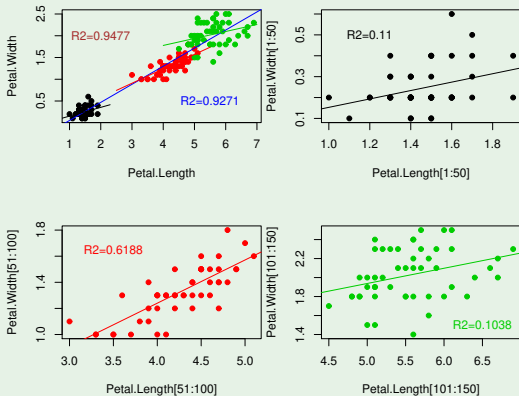
These are **almost** the same assumptions made in separate fits of each line, using only the  $n_i$  observations from each context.

But there are **stronger assumptions**: independence and variance homogeneity of random errors must be valid for the **pooled** data from all 3 contexts.

# A warning

## Mixing subpopulations may create illusions

Here are the scatterplots and values of  $R^2$  of Petal.Width vs. Petal.Length for: a **single line**, **ANCOVA** and separate fits, for the three iris species (*setosa*, *versicolor* and *virginica*).



# The $R^2$ of an ANCOVA model

It is possible to relate the Coefficients of Determination of the ANCOVA model,  $R^2$ , and of the  $k$  single-level models,  $R_{[i]}^2$ . We have:

$$R^2 = \frac{\sum_{i=1}^k R_{[i]}^2 SQT_i + SQF}{\sum_{i=1}^k SQT_i + SQF}.$$

where  $SQR_i$  and  $SQT_i$  are for observations from level  $i$ , and  $SQF$  is the Factor Sum of Squares in the one-way ANOVA of all observations, on the factor indicating the  $k$  contexts being compared (without the numerical predictor).

- if  $SQF \approx 0$  (i.e., the Factor has no significant effects on  $Y$ ),  $R^2$  is approximately a weighted mean of the  $R_{[i]}^2$  (with weights  $SQT_i$ ). In this case,  $R^2 \approx 1$  only if most  $R_{[i]}^2 \approx 1$ .
- for very large  $SQF$  (i.e., significant effects of the Factor on  $Y$ ), the differences in the means of  $Y$  for each group dominate the expression.  $SQF \gg \sum_{i=1}^k SQT_i \Rightarrow R^2 \approx 1$ , regardless of the  $R_{[i]}^2$ .

## Again the example on slide 357

The values of each Sum of Squares, and of the  $R^2$ , for each model mentioned on slide 357, are:

	SQT	SQR	SQRE	QMRE	R2
setosa	0.54420	0.05985029	0.4843497	0.01009062	0.1099785
versicolor	1.91620	1.18583401	0.7303660	0.01521596	0.6188467
virginica	3.69620	0.38349444	3.3127056	0.06901470	0.1037537
Ancova	86.56993	82.04251207	4.5274213	0.03144043	0.9477022

```
Result one-way ANOVA:      Petal.Width ~ Species
                        SQF=80.41333      SQRE=6.15660
```

The high value of the ANCOVA  $R^2$  essentially results from the large  $SQF$ .

**Warning:** the single regression line model for pooled data does not appear in this comparison.

# Generalizing

Extending these results to any  $k$ -level factor is immediate.

The basic **idea** used to compare regression lines in  $k$  different contexts **can be generalized to study any multiple linear regression in  $k$  different contexts.**

For each predictor, we allow the possibility of having additive effects for any coefficient (when compared to the coefficient of the first context). These additive effects may be different for each context.