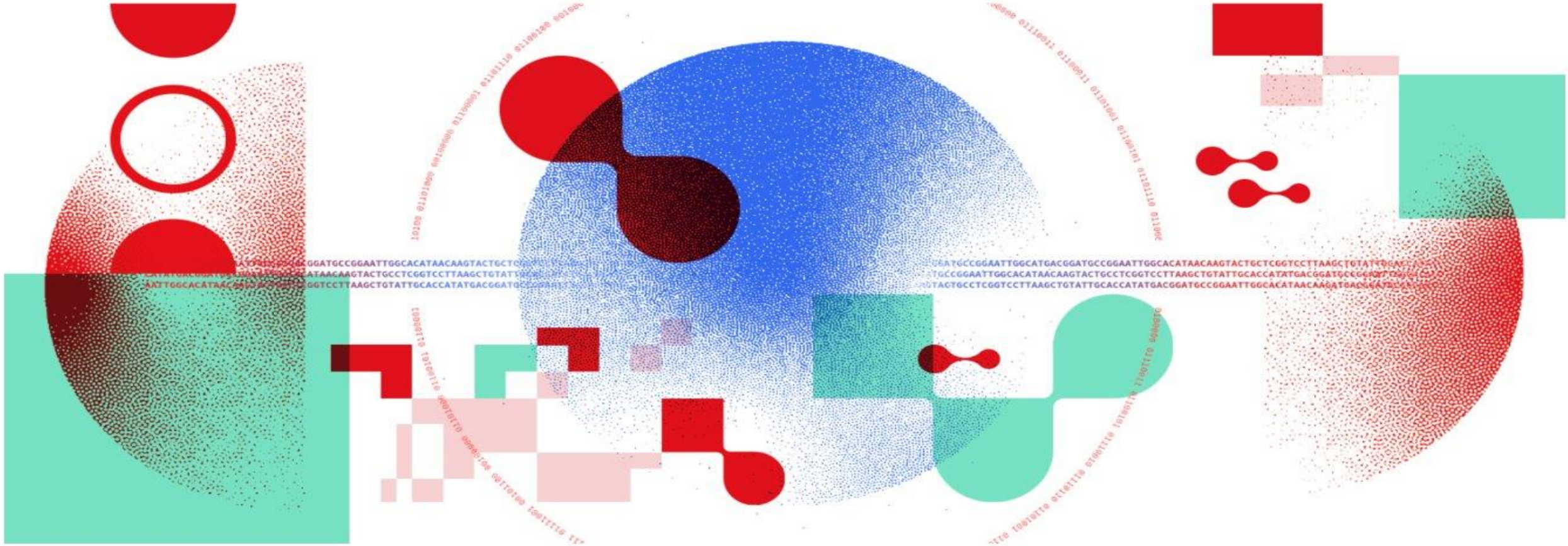


Advanced Statistical Modelling

Lausanne, August 2025

Joao Lourenço and Rachel Marcone

Simple and Multiple Linear Models

Plan

- Modelling- General rules
- Simple Linear Regression
- Multiple Linear Regression
- Assumptions
- Diagnostics

Modelling-General rules

- The idea of modelling is being able to find a good **representation** of the data in order maybe to understand what is **influencing** it or be able to **predict** future data.
- What is the best model ? Mostly it's a **trial and error**, and generally the patterns involved in the data are known (for example, linear, exponential, made of different steps).
- Start simple!
- Statistics help to understand if a model is better than another one but does not prove to be the only way to look at your data.
- Modelling is an approximation of the data and will never be perfect.

Statistical Modelling

- Statistical modelling is a set of **equations** that are solved, along with a set of random variables that follow certain **assumptions.**

- **Which equations ? Which assumptions ?**

Variables and questions

- **Dependent variables (or responses):** variables we want to describe, understand, explain, model, predict

- **Explanatory variables**
- **(or independent variables or predictors or covariates):** variables we use to explain, to describe or to predict the dependent variable(s). These are usually linked to the biological question you ask (for example : does gender have an effect on a drug response)

Modelling overview

- Many models will be written in the following formula
- $Y = f(X) + error$

- Sometimes a transformation of the Y variable is needed for a relationship so models can be also of the form
- $g(Y) = f(X) + error$

- Y is the dependent variable
- And here X is the only explanatory variables but there could be many
- $Y = f(X_1) + f(X_2) + f(X_3) + \ldots + error$
- The error term is what is of interest. We will try to **minimize** the error term and for that we will have to make assumptions on the error term.

Modelling overview in R

- The generic form of a model in R is given by
    - Response ~ predictions
- + to **add** more variables
- - to **leave out** variables
- : to indicate **interaction** between two terms
- * to include both interaction and the terms
- ^n adds all terms including interaction up to degree n
- I() treats what is in the parenthesis as mathematical expression

Modelling overview in R

- The generic form of a model in R is given by
    - Response ~ predictions
- + to **add** more variables
- - to **leave out** variables
- : to indicate **interaction** between two terms
- * to include both interaction and the terms
- ^n adds all terms including interaction up to degree n
- I() treats what is in the parenthesis as mathematical expression
    - **Income~ age^2 + I(weight/height^2) +gender**

Model formulas in R

- A simple formula in R can be for example

  - `yvar ~ xvar1 + xvar2 + xvar3`

- Can read ~ as “described (or modeled) by”.

- We could write a model (algebraically) as

  - $Y = b0 + b1\ x1 + b2\ x2 + b3\ x3 + error$

Model formulas in R

- By default, an intercept is included in the model – you don't have to include a term in the model formula

- If you want to leave the intercept out:
  - `yvar ~ -1 + xvar1 + xvar2 + xvar3`

- Can read ~ as "described (or modeled) by".

- We could write a model (algebraically) as
  - Y = b1 x1 + b2 x2 + b3 x3 + error

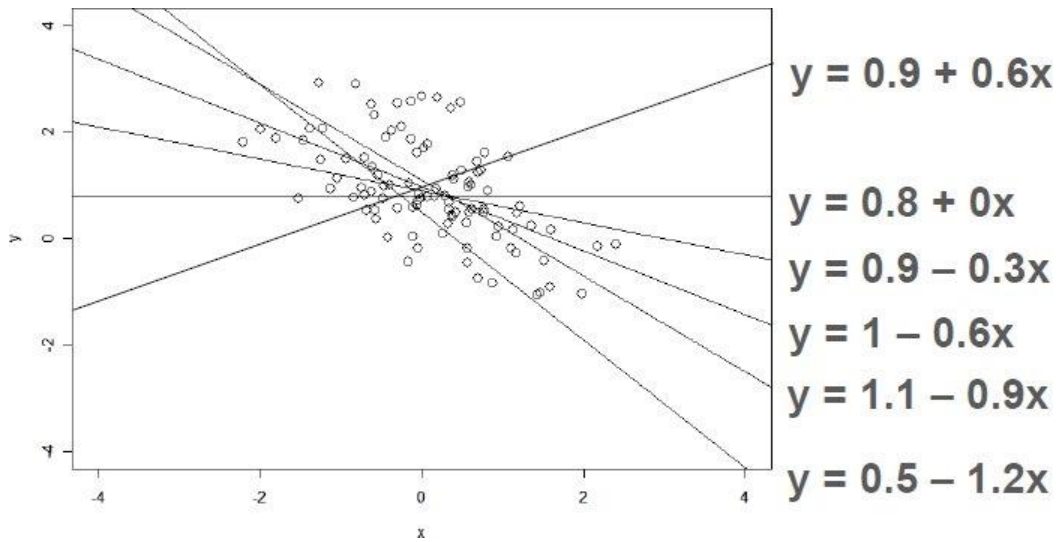
Model formulas in R

- If you only want the intercept, this is called the null model, or the estimation of the overall mean, or the grand mean
  - `yvar ~ 1`
- We could write a model (algebraically) as
  - Y = b0+ error

Linear regression

Linear Regression

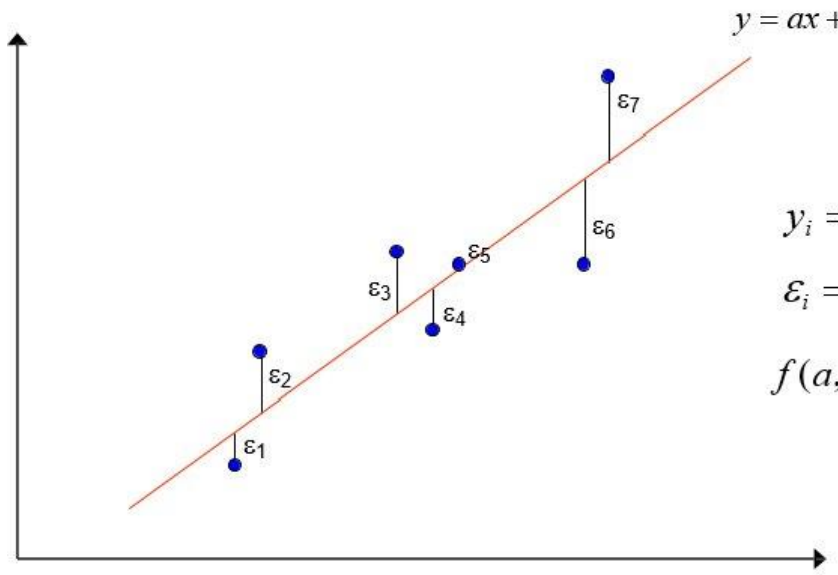
- Linear Regression = fit a line that summarises the relationship between x and y. Which is the best line ? What criteria ?



Least squares fitting (LS)

- One possibility-not the only one!
- The least square fitting finds the straight line with the smallest sum of squares of vertical errors

Least-square fitting



Regression line such that:

$$\sum_i \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + ..$$

minimum

$$y_i = ax_i + b + \varepsilon_i$$

$$\varepsilon_i = y_i - (ax_i + b)$$

$$f(a,b) = \sum_i \varepsilon_i^2 = \sum_i [y_i - (ax_i + b)]^2$$

$$\partial f(a,b)/\partial a = 0$$

$$\partial f(a,b)/\partial b = 0$$

Mathematically

Formalization and extension of linear regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$i = 1, \cdots, n$$

Y represents **one** data point

Y_i : response (known)

β_0, β_1 : model parameters (estimated)

X_i : predictor (known)

ε_i : error term $\sim N(0, \sigma^2)$ (estimated)

Minimizing $\sum_i \varepsilon_i^2$ yields b_0 and b_1 estimators of β_0 and β_1

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Error terms or Model residuals

- Distances between data points and the expected values based on the model (equation with fitted parameters)
- Model residuals represents the part of **variability** in the data the model was **unable to capture**
- Errors in least square fits are normally distributed with mean 0 and **constant variance**.

Warning

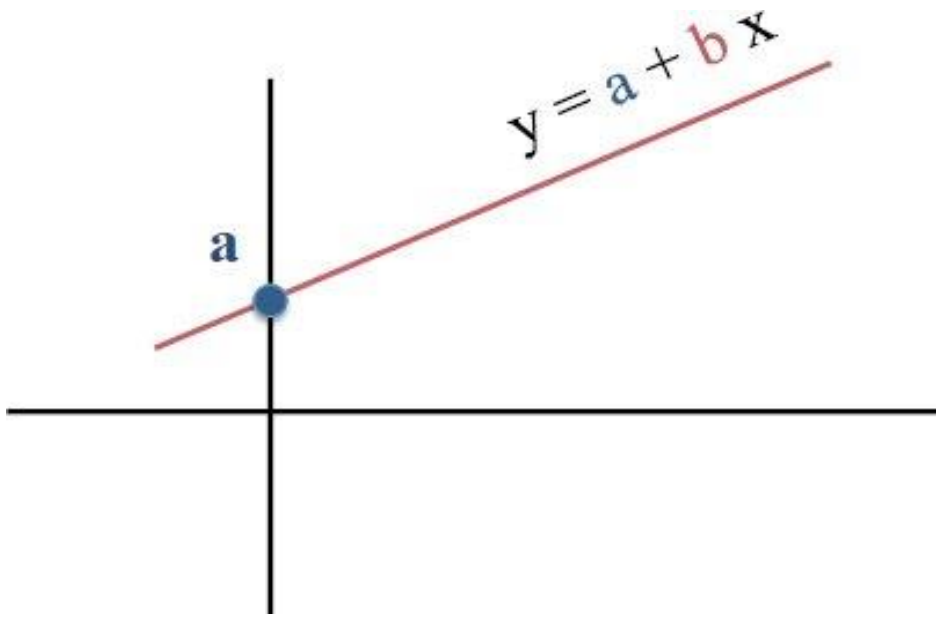
- It is always possible to fit a linear model and find a slope and intercept
- … but it does not mean that the model is meaningful

- Examination of the residuals is important
  - No trend in the residuals – any systemic effect should be captured by the model
  - Normality
  - Constant variance

Linear Regression

- The equation for a line to predict y knowing x is given by

- Y = a +bx where a is called the intercept and b the slope



Residuals

- The residuals are the error made when predicting the data using the regression line, therefore they are given by
    - Error = Ymeasured-Ypredicted = Ymeasured – (a+bX)
- The regression equation has the property that :
- The sum of the residuals is 0 which is the same as saying the mean of the residuals is 0

Multiple Linear regression

Categorical variable

Diagnostics

Hat Values and diagnostics

- $Y_{predicted} = X \beta$
- $X' Y = X' X \beta$ => because $X' (Y - X \beta) = 0$ as these planes are orthogonal
- $(X' X)^{-1} X' Y = \beta$
- $Y_{predicted} = X (X' X)^{-1} X' Y = H Y$

- This is the H matrix is the way to transform the measured values to the predicted values and is a very interesting matrix to look at.
- The diagonal values are a good indication of influence of the points in the regression

Hat values diagnostics

- Average value of h = number of coefficients (including the intercept)/number of observations= p/n
- In a simple case it would be 2/n

- A cutoff of 2*p/n or 3*p/n has been proposed in the litterature for influential points. *

*Belsley, Kuh, and Welsch (1980)

Linear models

Linearity is about the model parameters

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 \log X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$
$$Y_i = \beta \sin X_i + \varepsilon_i$$

} Linear in βs

$$Y_i = \beta_0 + \log(\beta_1 X_{i1} + \beta_2 X_{i2}) + \beta_3 X_{i3} + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 \exp(\beta_2 X_i + \beta_3) + \varepsilon_i$$

} Not linear in βs