Basic concepts illustrated with simple static models

David Makowski
INRA

Objectives

- Introduce several basic concepts using simple models.

- Give an introduction to the software R.

Outline

1. The different components of a model

2. A four-step approach for modelling

3. An introduction to R

4. Develop your own model with R

1. The different components of a model

Function

$$z = f(x, \theta)$$

Output variable

Input variables

Parameters

$$z = \theta_0 + \theta_1 x$$

- One input variable (*x*)
- One output variable (*z*)
- Two parameters (θ_0, θ_1)
- Linear function

The different components of a model

Component	Definition	Example
Output variable	Variable of interest computed by the model	Yield Disease incidence N_2O emission
Input variable	Variable measured before running the model and used to compute the output	Temperature Rain Soil depth
Parameter	Element used to compute the output, but not measured	Potential yield Coefficient of radiation interception
Function	Mathematical expression relating the output to the inputs and to the parameters	Linear Logistic Differential equation

Example 1

Models for predicting the occurence of high sclerotinia incidence in oilseed rape

Sclerotinia sclerotiorum, Lib., de Bary, in oilseed rape crops

- High variability of disease incidence across sites and years.
- High yield losses if disease incidence at harvest > 10%.
- Efficient chemical treatments exist, but are not always required.

Sowing Flowering Harvest *t*

n collected flowers.

↓

Incubation in Petri dishes

↓

m diseased flowers.

↓

% diseased flowers

Previous crops
Weather during the previous month
Soil characteristics

↓

Sum of risk points

Risk factor	Level	Points
Number of oil-seed crops during the last ten years	>5	30
	3-5	20
	2-3	10
	1	0
Other host crops during the last five years	Yes	15
	No	0
Level of infection in the last crop	High	15
	Moderate	5
	Low	0
Type of field	Wet	10
	Dry	0
Plant density	High	10
	Normal	5
	Low	0
Rain in the last month before flowering	More than normal	10
	Normal (50-60 mm)	5
	Less than normal	0

20 points

Example 1: models for predicting the occurence of high sclerotinia incidence in oilseed rape

Why a model?

- To predict the probability of high disease incidence at harvest (> 10%)
- To decide **at flowering** if a treatment is needed to control the disease.
- To avoid systematic fungicide application.

Example 1: models for predicting the occurence of high sclerotinia incidence in oilseed rape

% diseased flowers at flowering →

Sum of risk points at flowering →

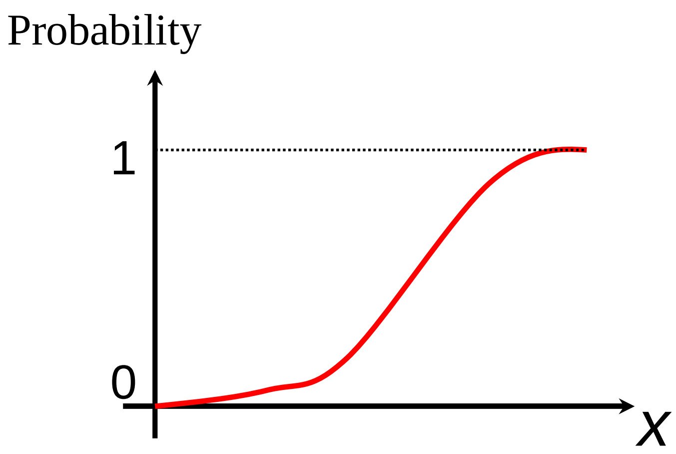
MODEL →

Probability that the disease incidence at harvest is higher than 10%

Example 1: models for predicting the occurence of high sclerotinia incidence in oilseed rape

Logistic model

$$\text{Probability of high disease incidence} = z = \frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)}$$



Example 1: models for predicting the occurence of high sclerotinia incidence in oilseed rape

Model 1: one input variable, x_1 = % diseased flowers

$$z = \frac{\exp\left(\theta_0 + \theta_1 x_1\right)}{1 + \exp\left(\theta_0 + \theta_1 x_1\right)}$$

Model 2: one input variable, x_2 = Sum of risk points

$$z = \frac{\exp\left(\theta_0 + \theta_2 x_2\right)}{1 + \exp\left(\theta_0 + \theta_2 x_2\right)}$$

Model 3: two input variables

$$z = \frac{\exp\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2\right)}{1 + \exp\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2\right)}$$

Example 1: models for predicting the occurence of high sclerotinia incidence in oilseed rape

Component	Model 1	Model 2	Model 3
Output variable	Proba. of high disease incidence	Proba. of high disease incidence	Proba. of high disease incidence
Input variable	% diseased flowers	Sum of risk points	% diseased flowers Risk points
Parameter	2	2	3
Function	Logistic	Logistic	Logistic

2. A four-step approach for modelling

i. Definition of input and output variables

ii. Definition of equations

iii. Parameter estimation

iv. Model evaluation

i. Definition of input and output variables

Variables are defined from

- the objectives of the modeller and model users

  e.g predicting disease incidence

- the knowledge about the system

  e.g effect of climatic variables on the development of a disease

- the availability of the data

  e.g weather station

→ A series of candidate variables

ii. Definition of the equations

- Equations must be defined to relate the output variables to the input variables

- Definition from the knowledge about the system

- Generally, several equations can be defined.

→A series of candidate model equations

e.g logistic equation

Step i + Step ii

↓

Series of candidate models based on different variables and/or different equations

3 models in the example

iii. Parameter estimation

- A model cannot run without parameter values !

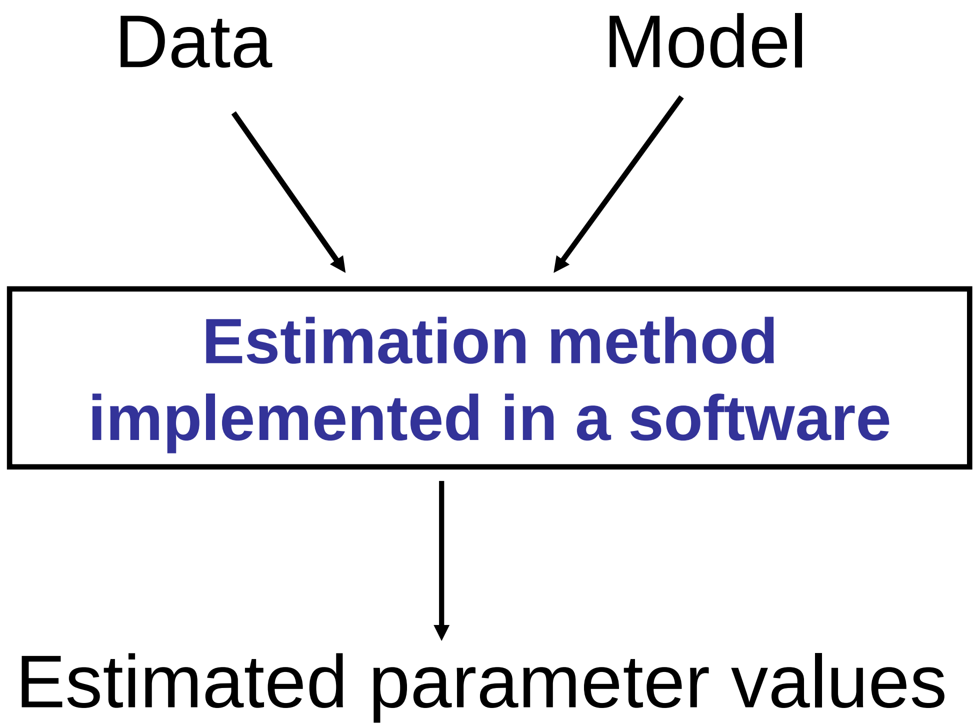
- Estimation = find values for the model parameters $(\theta_0, \theta_1, \theta_2)$

- Parameters can be estimated from data and expert knowledge

- Estimation requires **a method** and **a software** (R, Matlab…).

- The difficulty of this step depends on the complexity of the model

→ A set of parameter values for each candidate model


Data
Model
Estimation method
implemented in a software
Estimated parameter values

Estimation method

- An estimation method allows one to estimate model parameters from data

- Data = series of measured values of model inputs and outputs

- Several methods exist: **least squares**, **maximum likelihood** etc.

Estimation of the parameters of the logistic model 1 used for predicting the risk of sclerotinia

Logistic model 1: one input variable, two unknown parameters

$$z = \frac{\exp\left(\theta_0 + \theta_1 x_1\right)}{1 + \exp\left(\theta_0 + \theta_1 x_1\right)}$$

Estimation from 85 experimental plots by maximum likelihood

$$z = \frac{\exp\left(-3.61 + 5.36 x_1\right)}{1 + \exp\left(-3.61 + 5.36 x_1\right)}$$

85 plots in France

Diseased flowers	High incidence
0.06	1
0.09	0
0.83	1
0.13	0
0.03	0
...	...

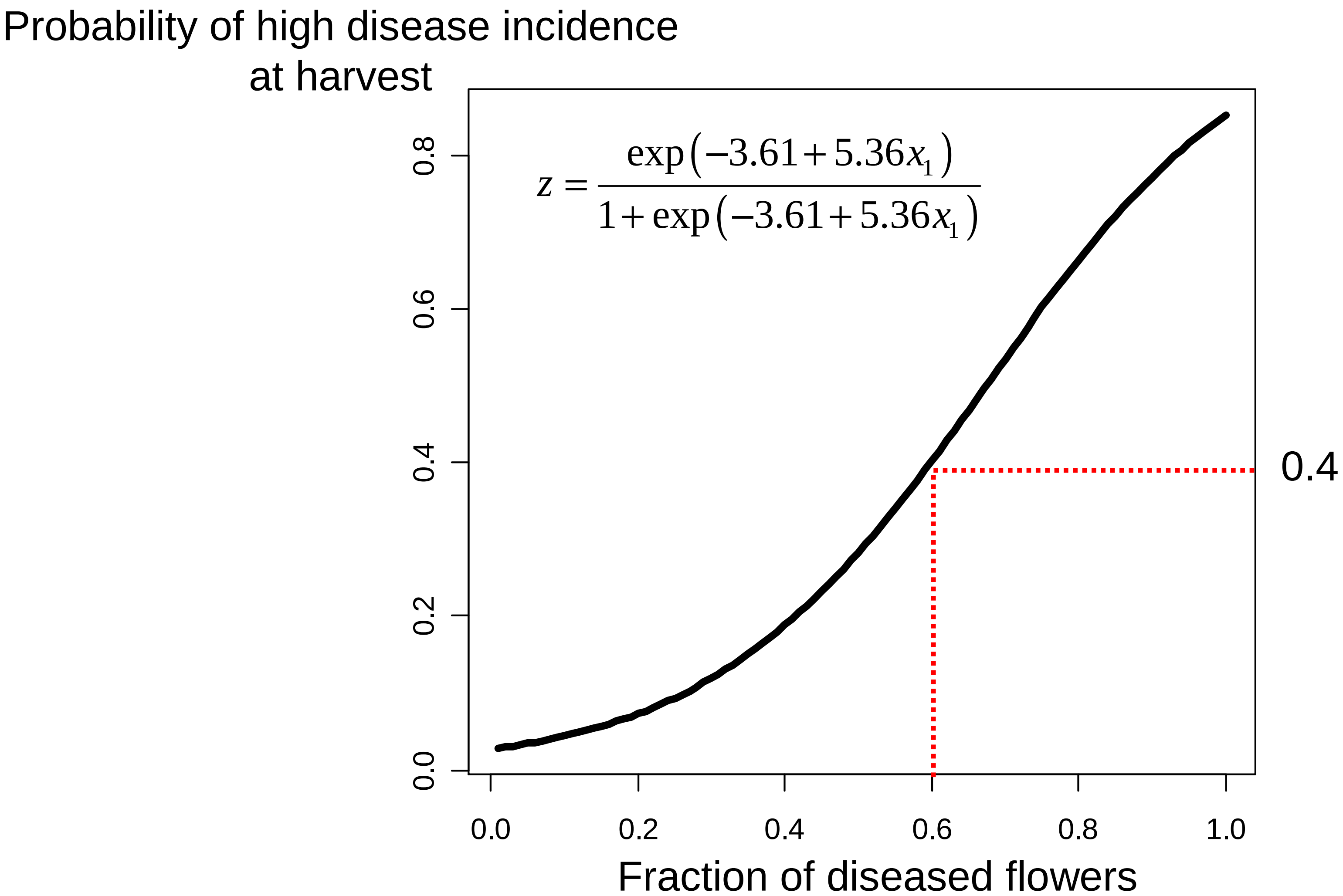
$$z=\frac{\exp\left(\theta_0+\theta_1 x_1\right)}{1+\exp\left(\theta_0+\theta_1 x_1\right)}$$

Maximum likelihood method implemented in a software

$$\hat{\theta}_0=-3.61,\ \hat{\theta}_1=5.36$$

$$z=\frac{\exp\left(-3.61+5.36x_1\right)}{1+\exp\left(-3.61+5.36x_1\right)}$$

Output from the logistic model 1 after estimation



iv. Model evaluation

- Each candidate model must be assessed using **one or several criteria.**

- Criteria must be chosen **in function of the objective** of the modeller and the model users.

- Data are generally required for this step, especially for assessing the accuracy of the model predictions.

→ **Model choice**

Steps iii and iv can be difficult

Specific lectures will be devoted to parameter estimation and model evaluation.

3. An introduction to R

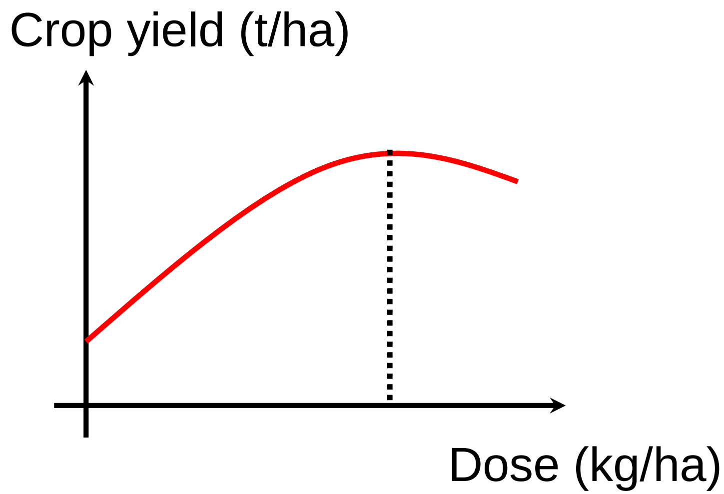
Example 2: models of wheat yield response to applied nitrogen fertilizer

Which nitrogen fertilizer dose should be applied ?

Example 2: models of wheat yield response to applied nitrogen fertilizer

Model 1

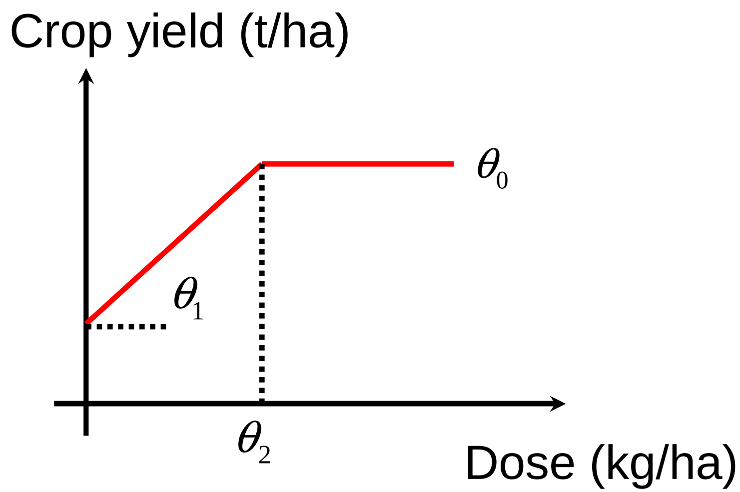
$$z = \theta_0 + \theta_1 x + \theta_2 x^2$$



Model 2

$$z = \theta_0 \text{ if } x \geq \theta_2$$

$$z = \theta_0 + \theta_1 (x - \theta_2) \text{ if } x < \theta_2$$



Example 2: models of wheat yield response to applied nitrogen fertilizer

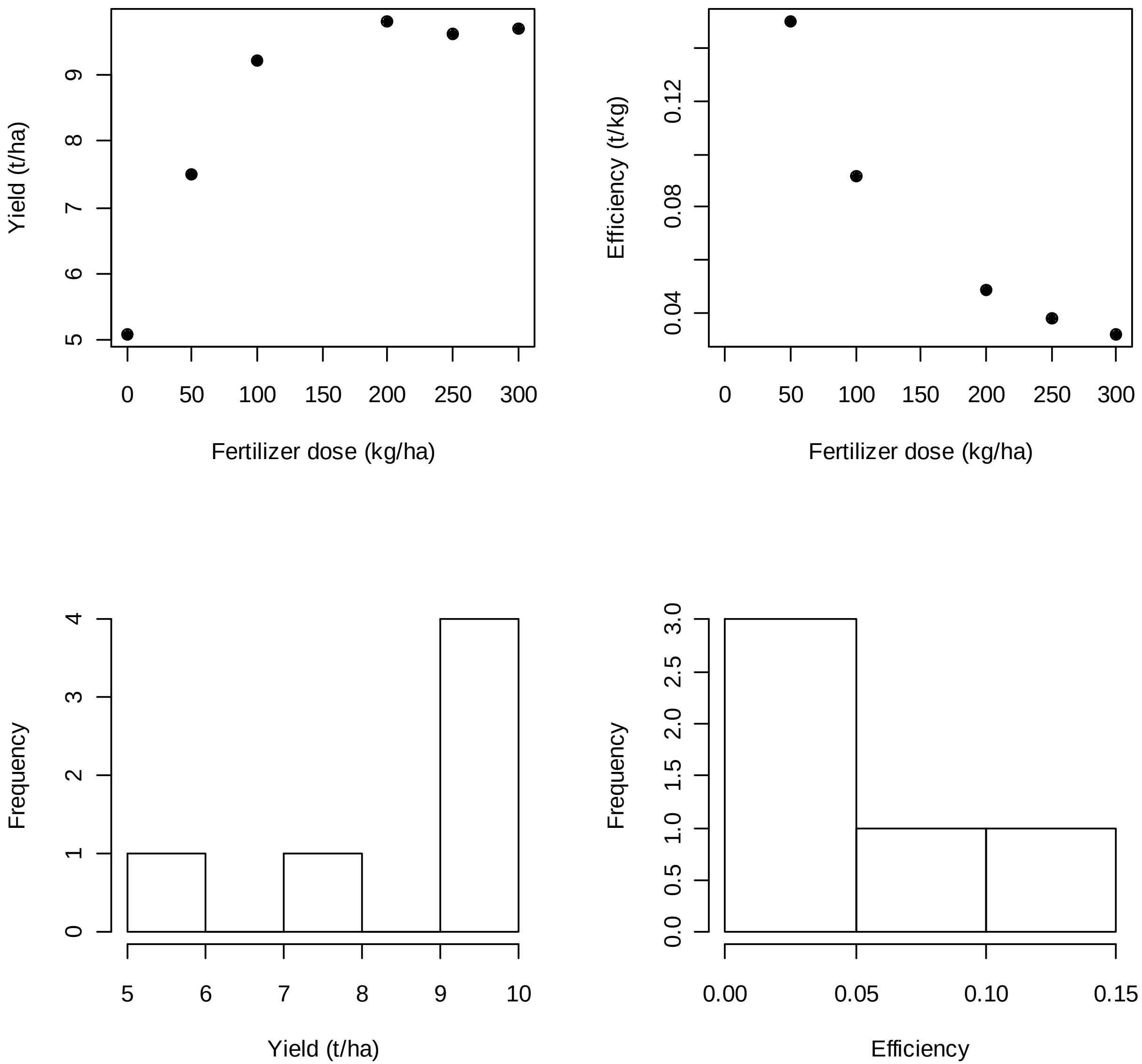
Component	**Model 1**	**Model 2**
Output variable	Yield (*z*)	Yield (*z*)
Input variable	Dose (*x*)	Dose (*x*)
Parameter	3 $(\theta_0, \theta_1, \theta_2)$	3 $(\theta_0, \theta_1, \theta_2)$
Function	Quadratic	Linear-plus-plateau

Reading data

```
x <- c(0, 50, 100, 200, 250, 300)       #Values of x (fertilizer dose)
y<- c(5.1, 7.5, 9.2, 9.8, 9.6, 9.7)     #Values of y (yield)
TAB<-data.frame(x, y)                   #dataset TAB
TAB                                     #print the dataset TAB
TAB$x                                   #column x of TAB
TAB$y                                   #column y of TAB
x                                       #column x of TAB
y                                       #column y of TAB
```

Graphics and data transformation

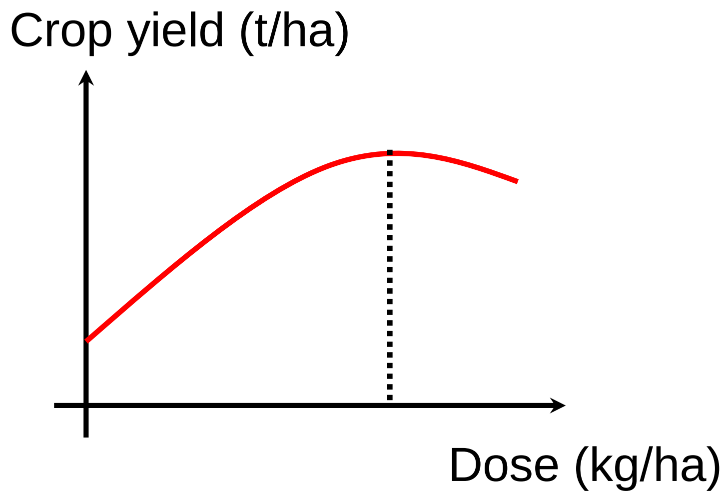
```
plot(x,y, xlab="Fertilizer dose (kg/ha)", ylab="Yield (t/ha)", pch=19)                    # plot y versus x
y<-y*1000                                                                                  # y in kg/ha
y<-y/1000                                                                                  # y in t/ha
Efficiency<-y/x                                                                            # new variable
plot(x, Efficiency, xlab="Fertilizer dose (kg/ha)", ylab="Efficiency (t/kg)", pch=19)      # new plot

X11()                                                                                      # new window
par(mfrow=c(2,2))                                                                          # split in four parts
plot(x,y, xlab="Fertilizer dose (kg/ha)", ylab="Yield (t/ha)", pch=19)                    # plot 1
plot(x, Efficiency, xlab="Fertilizer dose (kg/ha)", ylab="Efficiency (t/kg)", pch=19)      # plot 2
hist(y, xlab="Yield (t/ha)", main=" ")                                                     # plot 3
hist(Efficiency, main=" ")                                                                 # plot 4

mean(x)                                   # mean value of x
mean(y)                                   # mean value of y
mean(y[4:6])                              # mean of the last three values of y
summary(y)                                # distribution
length(y)                                 # length of y
```


Yield (t/ha)
Fertilizer dose (kg/ha)
Efficiency (t/kg)
Fertilizer dose (kg/ha)
Frequency
Yield (t/ha)
Frequency
Efficiency

Example 2: models of wheat yield response to applied nitrogen fertilizer

Model 1

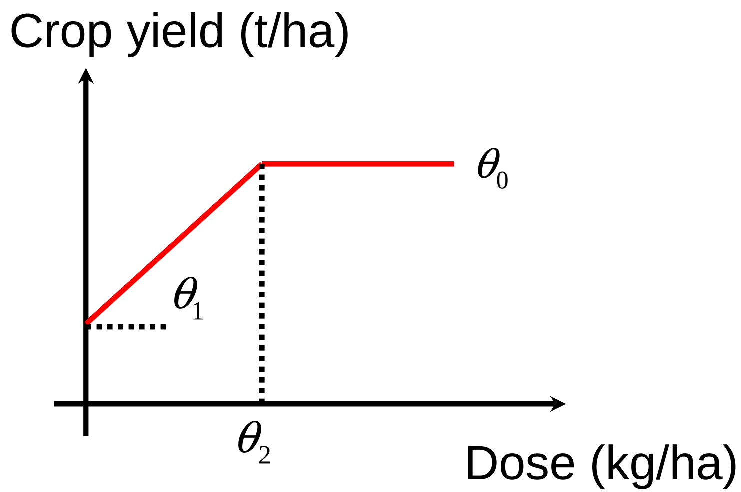
$$z = \theta_0 + \theta_1 x + \theta_2 x^2$$



Model 2

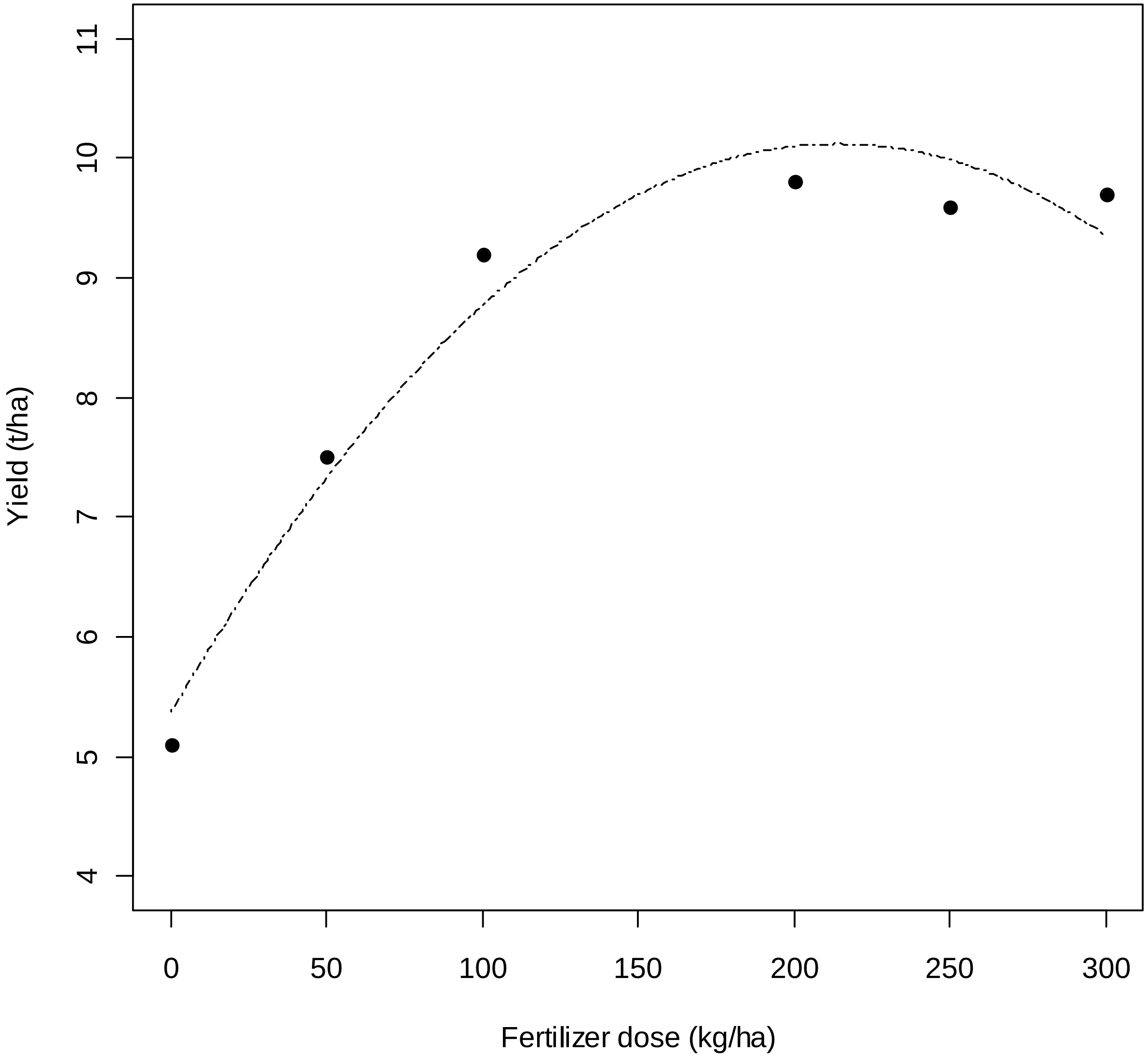
$$z = \theta_0 \text{ if } x \geq \theta_2$$

$$z = \theta_0 + \theta_1 \left(x - \theta_2 \right) \text{ if } x < \theta_2$$



Parameter estimation

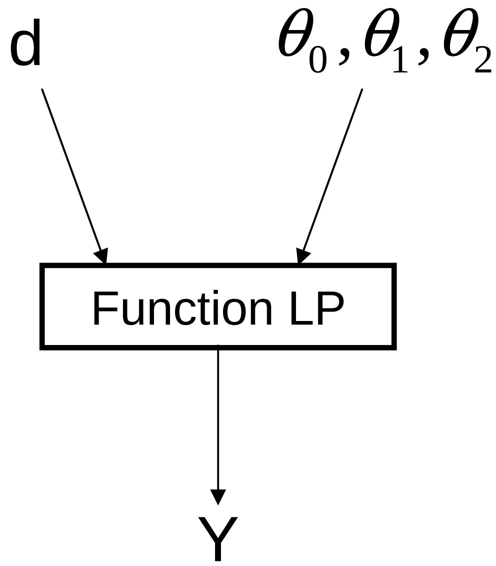
```
x2<-x*x                          # New variable
Fit<-lm(y~x+x2)                  # Parameter estimation by least squares
                                 # for the quadratic model

summary(Fit)                     # Results
coef(Fit)                        # The three estimated parameter values
Parameters<-coef(Fit)

X11()                            #New window
par(mfrow=c(1,1))

plot(x,y, xlab="Fertilizer dose (kg/ha)", ylab="Yield (t/ha)", pch=19, ylim=c(4,11))

Pred<-Parameters[1]+Parameters[2]*(0:300)+Parameters[3]*(0:300)^2

lines(0:300, Pred, lty=4)
```

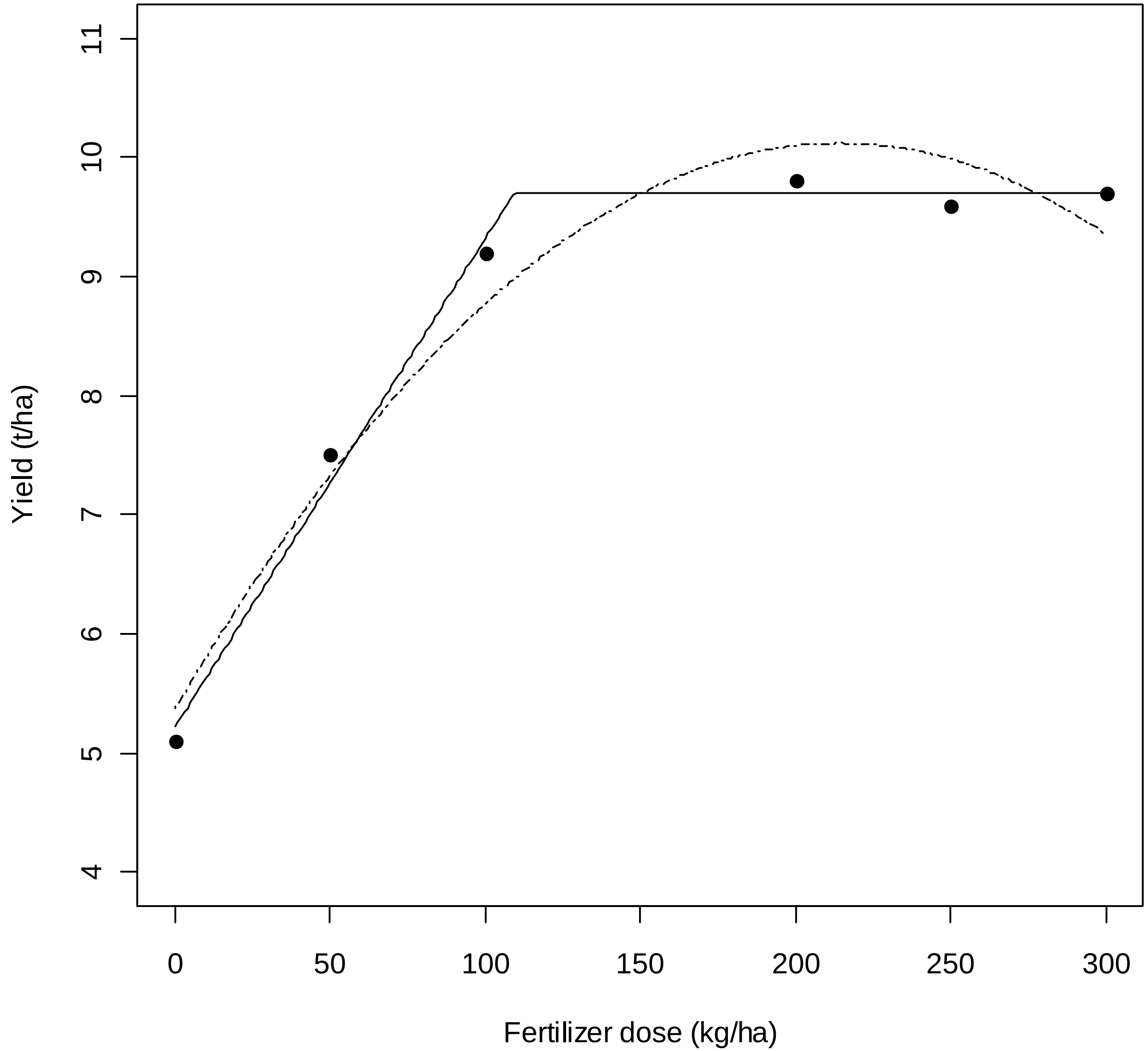


Yield (t/ha)
4
5
6
7
8
9
10
11
0
50
100
150
200
250
300
Fertilizer dose (kg/ha)

Parameter estimation

```
LP<-function(d, Theta0, Theta1, Theta2) {
        Y<-Theta0+Theta1*(d-Theta2)
        Y[d>=Theta2]<-Theta0
        return(Y)
        }

Fit<-nls(y~LP(x, Theta0, Theta1, Theta2), start=list(Theta0=9, Theta1=0.04,
Theta2=100), data=TAB)

summary(Fit)

Parameters<-coef(Fit)

Pred<-Parameters[1]+Parameters[2]*(0:300-Parameters[3])
Pred[Pred>Parameters[1]]<-Parameters[1]

lines(0:300, Pred)
```


d
$\theta_0, \theta_1, \theta_2$
Function LP
Y

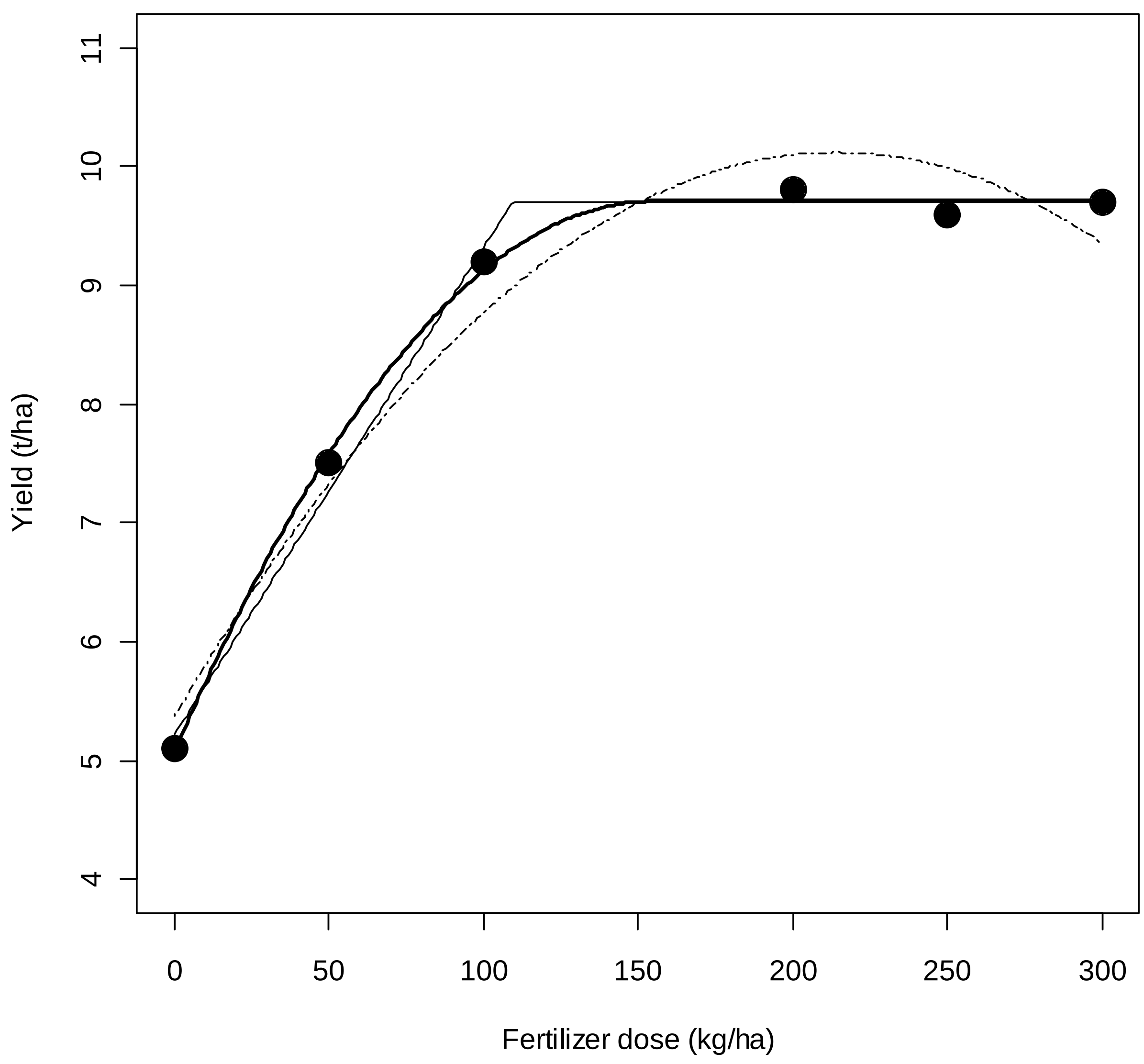


Yield (t/ha)
4
5
6
7
8
9
10
11
0
50
100
150
200
250
300
Fertilizer dose (kg/ha)

4. Build your own model

Develop a quadratic-plus-plateau model for relating yield to fertilizer dose

```
QP<-function(d, Theta0, Theta1, Theta2) {
            Y<-Theta0+Theta1*(d-Theta2)^2
            Y[d>=Theta2]<-Theta0
            return(Y)
            }

Fit<-nls(y~QP(x, Theta0, Theta1, Theta2), start=list(Theta0=9, Theta1=-
0.004, Theta2=100), data=TAB)

summary(Fit)

Parameters<-coef(Fit)

Pred<-Parameters[1]+Parameters[2]*(0:300-Parameters[3])^2
Pred[0:300>Parameters[3]]<-Parameters[1]

lines(0:300, Pred, lwd=2)
```



Yield (t/ha)
4
5
6
7
8
9
10
11
0
50
100
150
200
250
300
Fertilizer dose (kg/ha)

References related to the examples

Ennaïfar, S., D. Makowski, J-M. Meynard, Ph. Lucas. 2007. Evaluation of models to predict take-all incidence on winter wheat as a function of cropping practices, soil, and climate. *European Journal of Plant Pathology* 118:127-143.

Makowski, D., M. Taverne, J. Bolomier, M. Ducarne. 2005. Comparison of risk indicators for sclerotinia control in oilseed rape. *Crop Protection* 24:527-531

Makowski, D., M. Lavielle. 2006. Using SAEM (stochastic approximation of EM) to estimate parameters of models of response to applied fertilizer. *Journal of Agricultural, Biological, and Environmental Statistics* 11 (1):45-60.

Primot, S., M. Valantin-Morison, D. Makowski. 2006. Predicting the risk of weed infestation in winter oilseed rape crops. *Weed Research* 46:22-33.