An introduction to modelling, Poznan, Nov. 2008

Estimation of model parameters

David Makowski

INRA

Parameters

$$f\left(x,\theta\right)$$

Equations | Input variables | **Parameters**

« **A parameter** is a numerical value which is not calculated by the model and is not measured »

Parameter estimation

« aims at approximating the parameter values by using *experimental data* and/or *expert knowledge* »

It is important because

« *Model performances* depend on the accuracy of the parameter estimates »

The Bayesians and the Frequentists

For Frequentists,

- **parameters are fixed**
- **parameters are estimated by points values for a given dataset**
- **estimation is performed by using data only**

For Bayesians,

- **parameters are defined as random variables**
- **parameters are estimated by distributions for a given dataset**
- **estimation is performed from both data and prior information**
- **computations more complex, but results more intuitive**

Four steps for estimating parameters

1. How many and which parameters should be estimated?

→ In simple models, all parameters

→ In complex models, a subset of parameters is estimated

2. What kind of information is available?

→ Data

→ Prior information (expert knowledge, litterature)

3. Which estimation method?

→ Ordinary least squares

→ Weighted/Generalized least squares, maximum likelihood

→ Bayesian method

4. What is the accuracy of the parameter estimator?

→ Theoretical consideration, variances, residuals

Four estimation problems

Pb.A: One parameter

Pb.B: Linear with 2 parameters

Pb.C: Non linear with 18 parameters

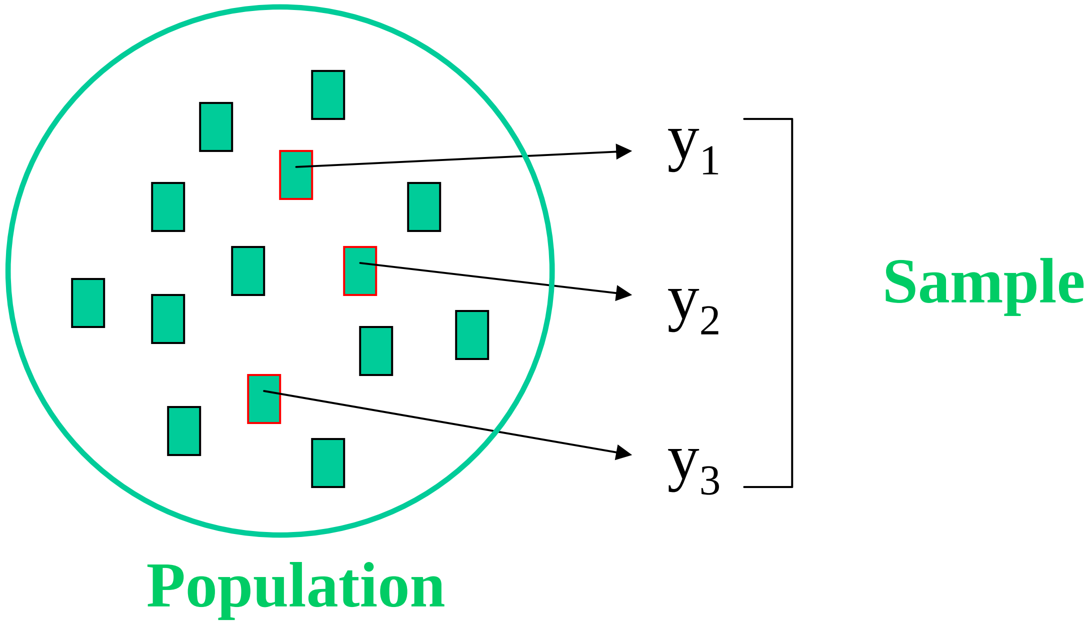
Estimation from data only

Pb.D: Estimation from data and prior information

Problem A

« Estimation of the average oilseed rape yield in 2004 in a small area

from 3 yield measurements collected in three different plots »



Step 1. Which parameters?

A single parameter, the average yield in the considered area, noted θ.

Step 2. What kind of information?

Available information: a ***sample*** of three measures collected in three plots from the ***population of plots*** of interest

Step 3. Which method?

An estimator of the average yield is:

$$\hat{\theta} = \frac{y_1 + y_2 + y_3}{3}$$

Example :

- If y_1=30, y_2=39 et y_3=35, the estimated average yield is **34.7** q/ha.
- If y_1=32, y_2=38 et y_3=39, the estimated average yield is **36.3** q/ha.

« An estimator is a function relating the parameter to the observations »

Data set 1 $\longrightarrow$ $\hat{\theta}_1$

Data set 2 $\longrightarrow$ $\hat{\theta}_2$

Data set N $\longrightarrow$ $\hat{\theta}_N$

Step 4. Is the estimator accurate?

$$E\left[\left(\hat{\theta}-\theta\right)^2\right]=\left[E\left(\hat{\theta}\right)-\theta\right]^2+\mathrm{var}\left(\hat{\theta}\right)$$

Mean squared error — **Bias^2** — **Variance**

Step 4. Is the estimator accurate?

a. Theoretical consideration

« Under some assumptions, our estimator is ***unbiased*** and of ***minimum variance*** among the unbiaised estimators »

Step 4. Is the estimator accurate?

b. Estimator variance

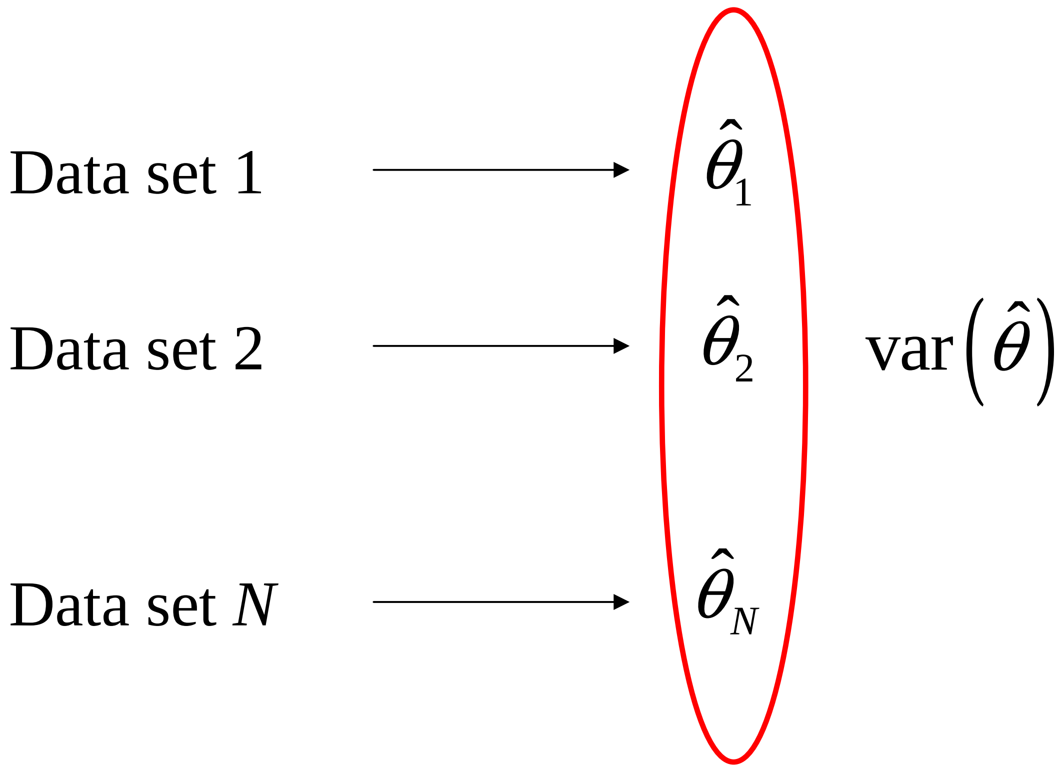
$\mathrm{var}\left(\hat{\theta}\right)$ can be estimated from data

Example :

• If y_1=30, y_2=39 and y_3=35, the estimated variance is 6.78 q²/ha², standard deviation=2.6 q/ha.

• If y_1=32, y_2=38 and y_3=39, the estimated variance is 4.78 q²/ha², standard deviation=2.19 q/ha.

The variance of an estimator measures its variability across datasets



Problem B

« Estimation of the parameters of the model $f(x;\theta_1,\theta_2)$ »

$$f(x;\theta_1,\theta_2) = \theta_1 + \theta_2\, x$$

Nitrogen uptake in oilseed rape crop

Nitrogen fertilizer dose

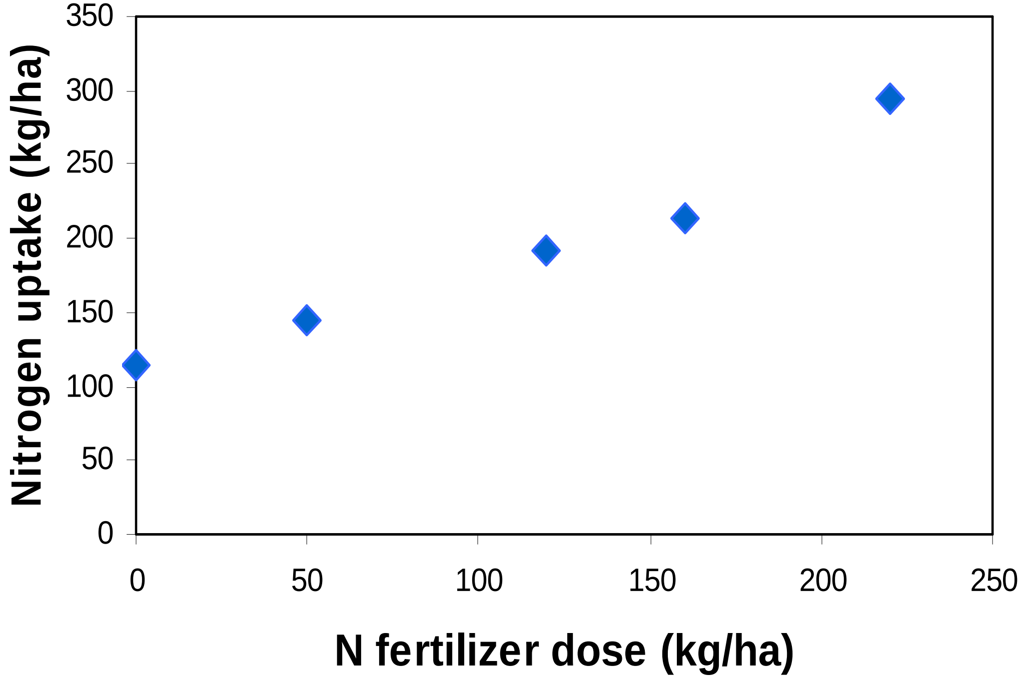
This model computes nitrogen uptake in function of fertilizer dose

Step 1. Which parameters?

Two model parameters: θ_1 and θ_2

Step 2. What kind of information?

A ***sample*** of 5 nitrogen uptake measurements obtained in 5 plots in the ***population of interest*** (an area in France)



Step 3. Which method?

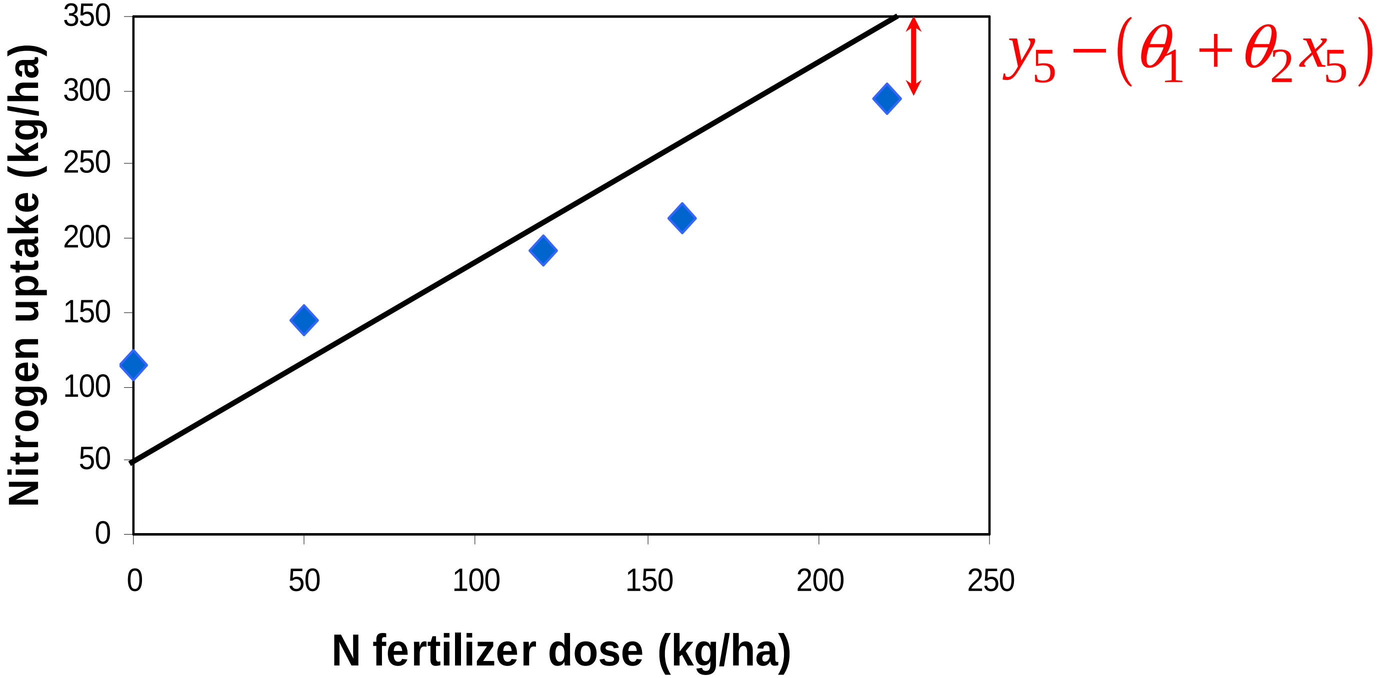
Ordinary least squares

The parameter estimators are the values of θ_1 and θ_2 minimizing

$$\sum_{i=1}^{N}\left(y_i-\theta_1-\theta_2 x_i\right)^2$$

that is

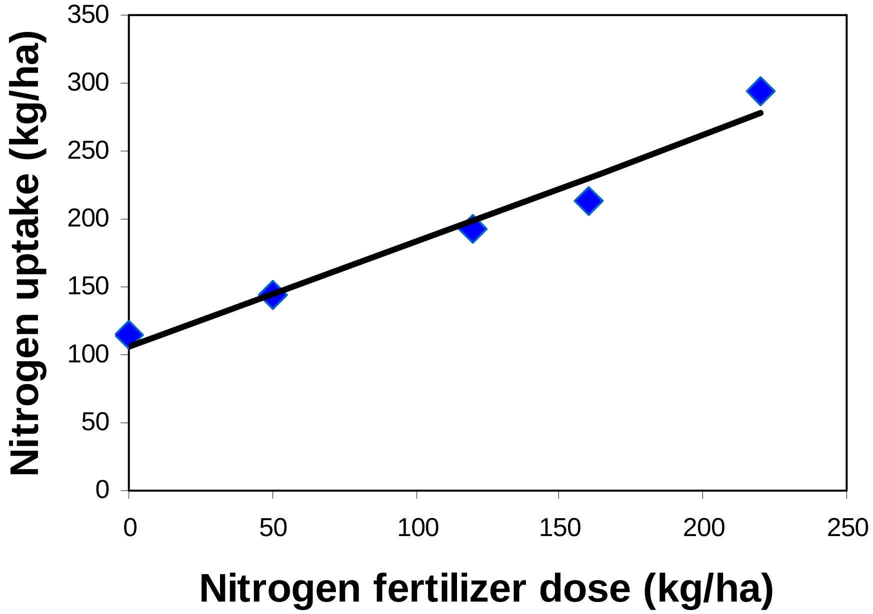
$$\hat{\theta}_2=\frac{\sum_{i=1}^{N}\left(y_i-\bar{Y}.\right)\left(x_i-\bar{X}.\right)}{\sum_{i=1}^{N}\left(x_i-\bar{X}.\right)^2} \qquad \hat{\theta}_1=\bar{Y}.-\hat{\theta}_2\,\bar{X}.$$



350
300
250
200
150
100
50
0
0
50
100
150
200
250
Nitrogen uptake (kg/ha)
N fertilizer dose (kg/ha)
$y_5 - (\theta_1 + \theta_2 x_5)$

Here, with our 5 measurements, we got $\hat{\theta}_1 = 106.01 \text{ kg.ha}^{-1}$

$$\hat{\theta}_2 = 0.78 \text{ kg.kg}^{-1}$$



Step 4. Are these estimators accurate?

$$E\left[\left(\hat{\theta}-\theta\right)^2\right]=\left[E\left(\hat{\theta}\right)-\theta\right]^2+\operatorname{var}\left(\hat{\theta}\right)$$

Mean squared error — **Bias²** — **Variance**

Step 4. Are these estimators accurate?

a. Theoretical aspect

« Under some assumptions,these estimators are ***unbiased*** and with ***minimum variances*** among the unbiased estimators ».

Assumptions are:

- ***independance*** of the model errors,
- ***homogeneity*** of the model error variances.

Step 4. Are these estimators accurate?

b. Variances of the estimators

Estimation of $\mathrm{var}\left(\hat{\theta}\right)$ from the data

$$\sqrt{\mathrm{var}\left(\hat{\theta}_1\right)} = 11.99 \ \ \mathrm{kg.ha}^{-1}$$

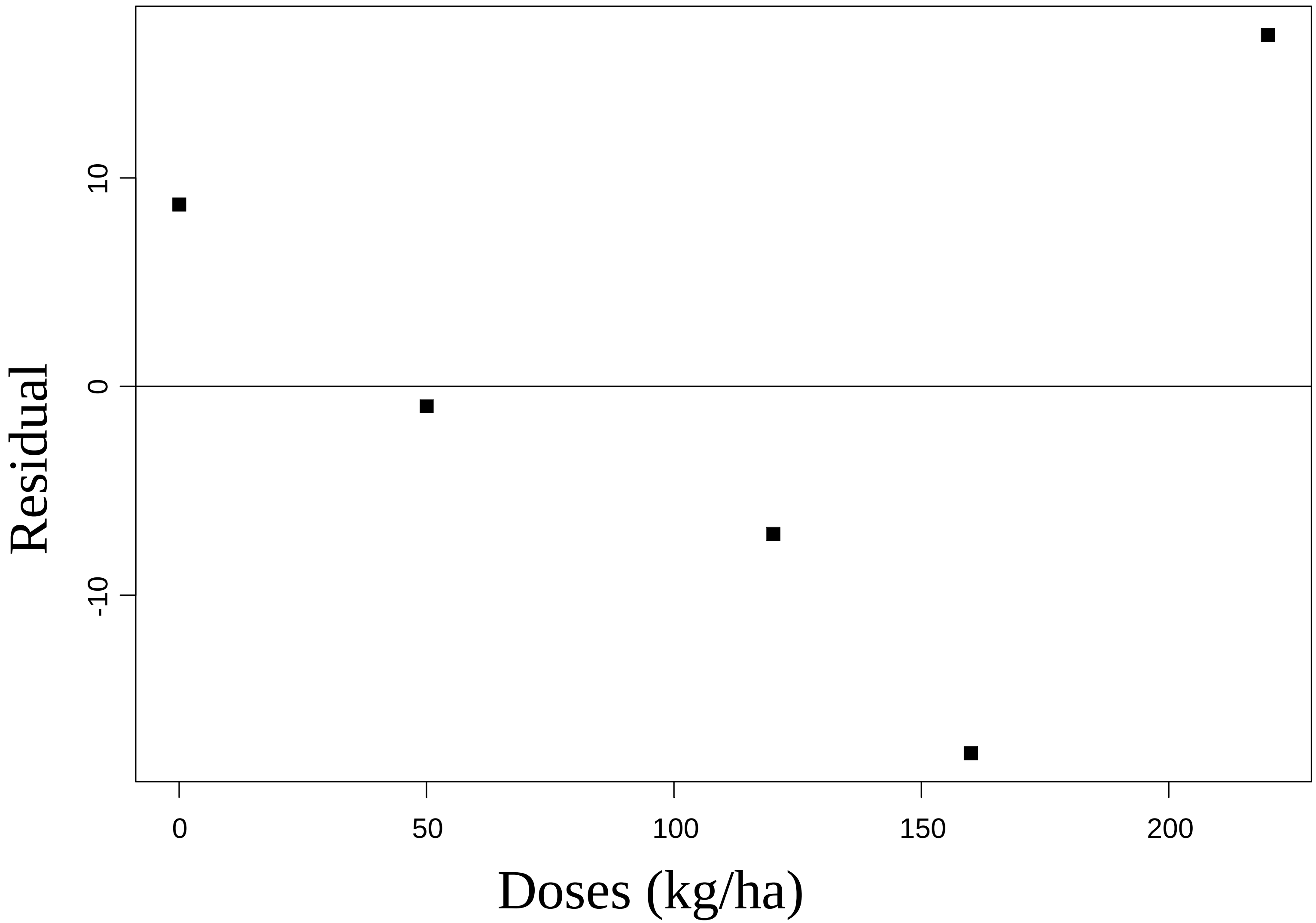
$$\sqrt{\mathrm{var}\left(\hat{\theta}_2\right)} = 0.09 \ \ \mathrm{kg.kg}^{-1}$$

Step 4. Are these estimators accurate?

c. Analysis of the residuals

$$r_i = y_i - \left(\hat{\theta}_1 + \hat{\theta}_2 x_i\right), \quad i = 1, \ldots, 5$$

Useful to check the independance of the model errors and variance homogeneity



Residual
10
0
-10
0
50
100
150
200
Doses (kg/ha)

R code

```
DOSE<-c(0,50,120,160,220)
NABS<-c(114.75,144.0,192.38,213,294.16)
DATA<-data.frame(DOSE,NABS)

Fit<-lm(NABS~DOSE,data=DATA)

print(summary(Fit))
plot(DOSE,Fit$residuals,ylab="Residual",ylab="Dose",pch=15)
abline(0,0)
```

Comments on the first two problems

Four steps

1. **Which parameters?**

2. **What kind of information?**

3. **Which estimation method?**

4. **Accuracy of the parameter estimators?**

Comments on the first two problems

It was easy because

- Linear model: $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$

  → Analytic relationships between estimators and data

- Number of data > Number of parameters

- Only one type of measurements

- No prior information about parameter values

- Sotwares are available for the computations (SAS, R, ModelMaker...).

It can be much more difficult

- Non linear models: $\neq \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$

  → No analytic relationship between the estimators and the data

- The number of data can be low compared to the number of parameters

- Complex dataset

  → several types of observations, correlated observations

- Prior information about parameter values

A much more difficult problem !

→ Non linear model

→ Many parameters

→ Prior information

→ Several types of measurements collected in several plots

Problem C

Estimation of the parameters of a model simulating winter wheat growth between January and May

(Jeuffroy et Recous, 1999)

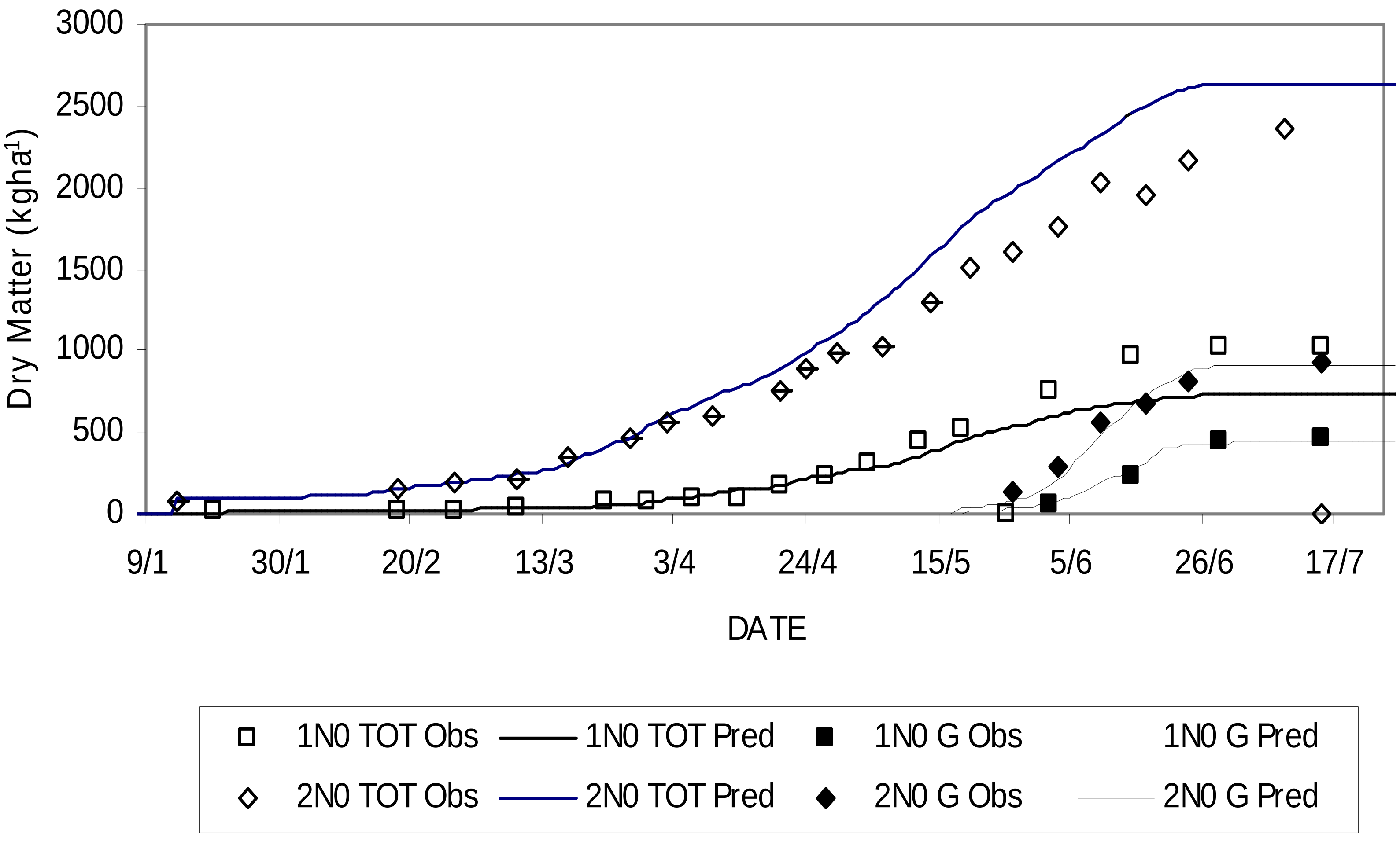
State variables simulated at a daily time step

- Crop above-ground biomass (dry matter) (kg/ha) → MS_t
- Nitrogen uptake (kg/ha), → QN_t
- Leaf Area Index → LAI_t

Input variables:

- Global daily radiation → RG_t
- Average daily temperature, → T_t
- Initial values of biomass and nitrogen uptake → MS_0, QN_0

Simulations of wheat biomass using the AZODYN dynamic crop model



Problem C

Few model equations

$$MS_j = MS_{j-1} + \left(E_{b\max} \times ft_{j-1} \times Ei_{j-1} \times C \times RG_{j-1} \right)$$

$$Ei_{j-1} = E_{i\max} \left[1 - \exp\left(-K \times LAI_{j-1} \right) \right]$$

$$LAI_{j-1} = D \times QNc_{j-1}$$

$$MS_j = MS_{j-1} + \left\{ E_{b\max} \times C \times E_{i\max} \left[1 - \exp\left(-K \times D \times QNc_{j-1} \right) \right] \times ft_{j-1} \times RG_{j-1} \right\}$$

18 parameters

Parameter	Meaning	Initial value	Range
Ebmax	Efficiency of radiation conversion	3.3 g/MJ	1.8-4
K	Coefficient of radiation extinction	0.72	0.6-0.8
D	Ratio LAI / Critical nitrogen uptake	0.028	0.02-0.045
Vmax	Maximum rate of nitrogen uptake	0.5 kg/ha/dj	0.2-0.7
C	PAR/RG	0.48	
Tmin	Minimum temperature for photosynthesis	0 °C	
Topt	Optimum temperature for photosynthesis	15 °C	
Tmax	Maximum température for photosynthesis	40 °C	
Eimax	Efficiency of radiation interception	0.96	
Tep-flo	Time between two stages	150 dj	
E		1.55 t/ha	
F		4.4 %	
G		5.35 %	
H		-0.442	
L		2 t/ha	
M		6 %	
N		8.3 %	
P		-0.44	

The two expressions of a dynamic model

1: Dynamic system model

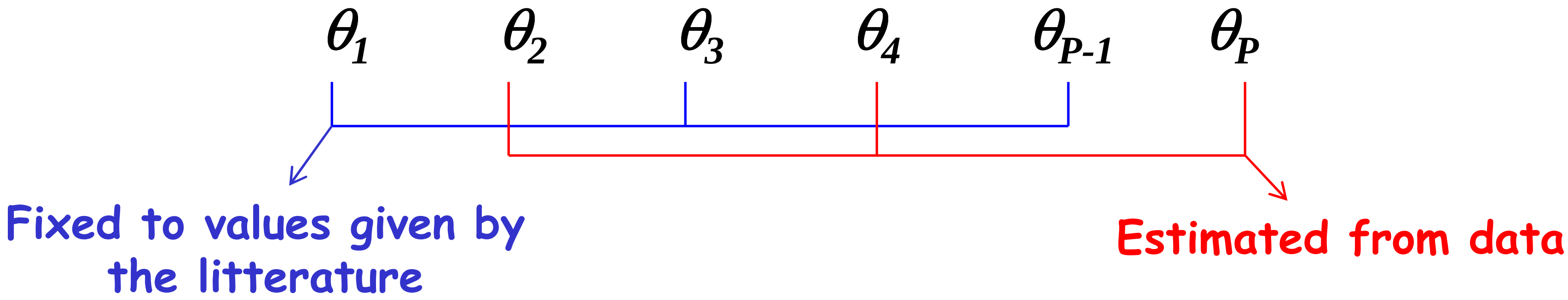
$$MS_t = MS_{t-1} + g\left(X_{t-1};\theta\right)$$

2: Response model

$$MS_t = f\left(t, X;\theta\right)$$

Step 1. Which parameters?

A subset of parameters must be selected



- **Numerical problems if all the parameters are estimated**

- **Not a good idea anyway**

  → High estimator variances

  → High prediction errors

New issue: How to select the subset of parameters

i. **from the literature**

ii. **by analyzing the model equations**

iii. **by sensitivity analysis**

iv. **from data**

i. from the literature

« Determine the parameters whose values are not well known ».

Drawbacks :

- can be quite subjective
- the available papers are not always relevant

ii. by analyzing the model equations

« Identify the parameters which cannot be simultaneously estimated »

$$MS_j = MS_{j-1} + \left(E_{b\max} \times ft_{j-1} \times Ei_{j-1} \times C \times RG_{j-1} \right)$$

$$Ei_{j-1} = E_{i\max} \left[1 - \exp\left(-K \times LAI_{j-1} \right) \right]$$

$$LAI_{j-1} = D \times QNc_{j-1}$$

$$MS_j = MS_{j-1} + \left\{ E_{b\max} \times C \times E_{i\max} \left[1 - \exp\left(-K \times D \times QNc_{j-1} \right) \right] \times ft_{j-1} \times RG_{j-1} \right\}$$

Case 1: only observed values of MS are available

Case 2: observed values of MS and LAI are available

$$MS_j = MS_{j-1} + \left(E_{b\max} \times ft_{j-1} \times Ei_{j-1} \times C \times RG_{j-1} \right)$$

$$Ei_{j-1} = E_{i\max} \left[1 - \exp\left(-K \times LAI_{j-1} \right) \right]$$

$$LAI_{j-1} = D \times QNc_{j-1}$$

$$MS_j = MS_{j-1} + \left\{ E_{b\max} \times C \times E_{i\max} \left[1 - \exp\left(-K \times D \times QNc_{j-1} \right) \right] \times ft_{j-1} \times RG_{j-1} \right\}$$

Case 1: only observed values of MS are available

It is not possible to estimate simultaneously
- the 3 parameters $E_{b\max}, C, E_{i\max}$
- the 2 parameters K, D

Case 2: observed values of MS and LAI are available

It is not possible to estimate simultaneously the 3 parameters $E_{b\max}, C, E_{i\max}$

iii. by sensitivity analysis

« Select the parameters that strongly influence the model outputs »

Drawbacks:

A sensitivity threshold must be defined.

Does not prevent from lack of identifiability .

iv. from data

« Select the parameters leading to the best model predictions »

Number of estimated parameters	MSEPcv
1	$MSEP_1$
2	$MSEP_2$
3	$MSEP_3$
…	…
P	$MSEP_P$

Step 1. Which parameters?

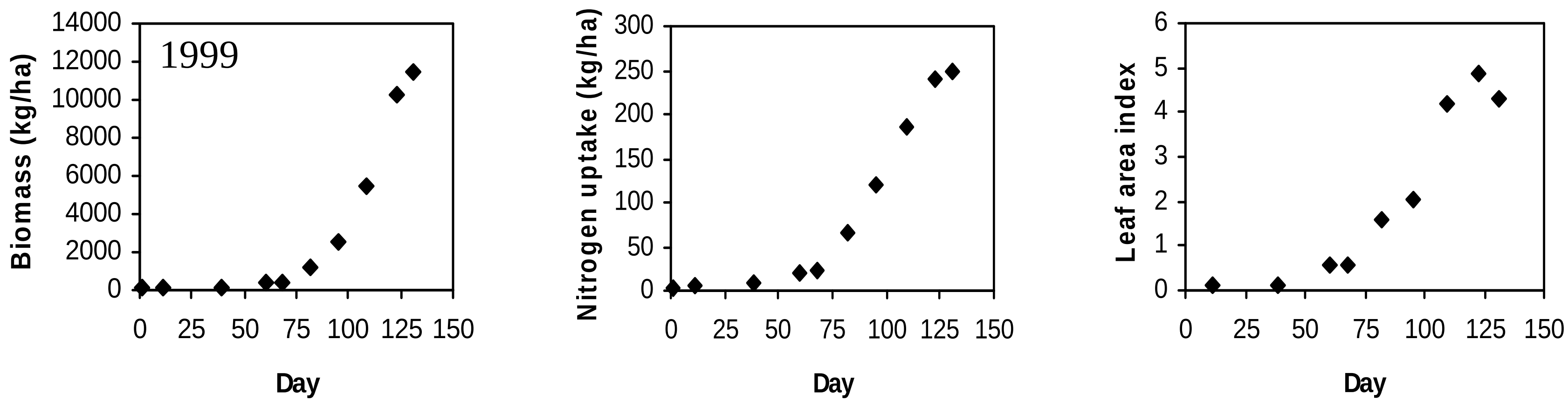
- **13 parameters** were fixed to values provided by the literature.
- **One parameter** was fixed after an analysis of the model equation.
- **Four parameters** were estimated from data: E_{BMAX}, D, K and V_{MAX}

Step 2. What kind of information?

• Measurements of wheat **biomass**, of **leaf area index** and of **nitrogen uptake** for one site (Grignon) and 6 years.

• Ten dates of measurement each year.

• Three replicates at each date. Replicates were averaged.



Step 3. Which estimation method?

1st option: Ordinary least squares

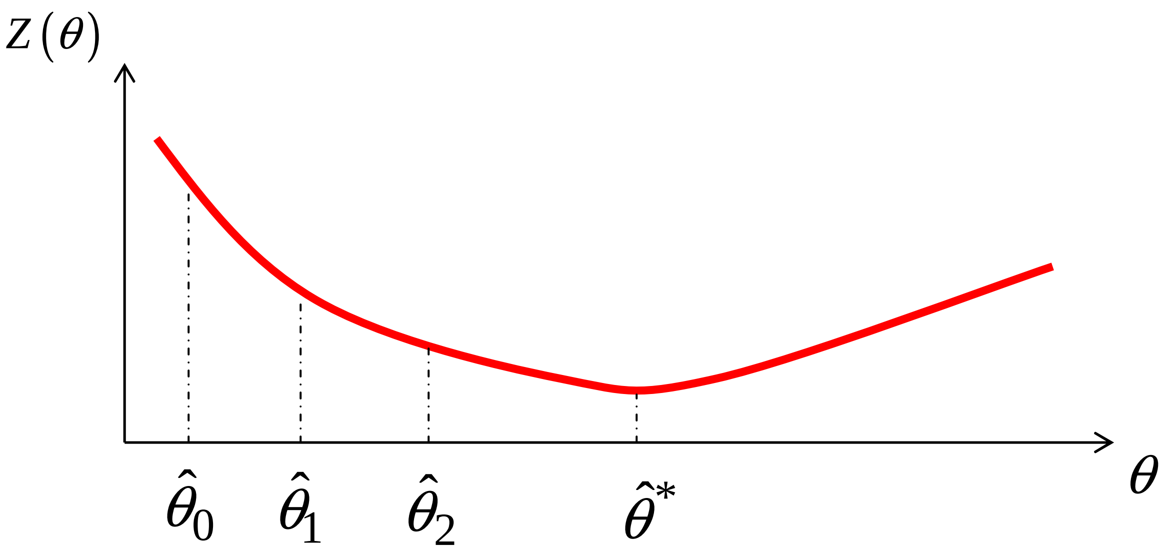
Find θ minimizing: $$Z(\theta) = \sum_{i=1}^{N} \left[y_i - f(t_i, x_i; \theta) \right]^2$$

Difficulties:

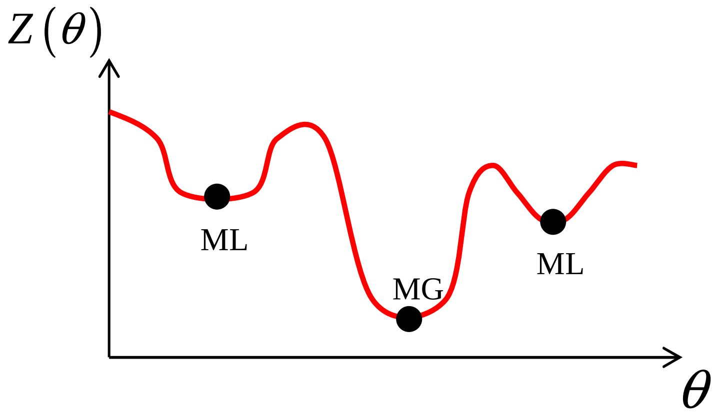
- non linear model

- no analytical expression for the estimators

Minimization using an iterative algorithm



Local optimum, global optimum



→ Try several starting values !

Practical considerations

• Several programs were developed to implement this kind of algorithm (SAS, R, MatLab, Fortran, C++…)

• They use the following entries:

- data
- a model equation,
- initial parameter values.

• The output is a set of estimated parameter values.

Step 3. Which estimation method?

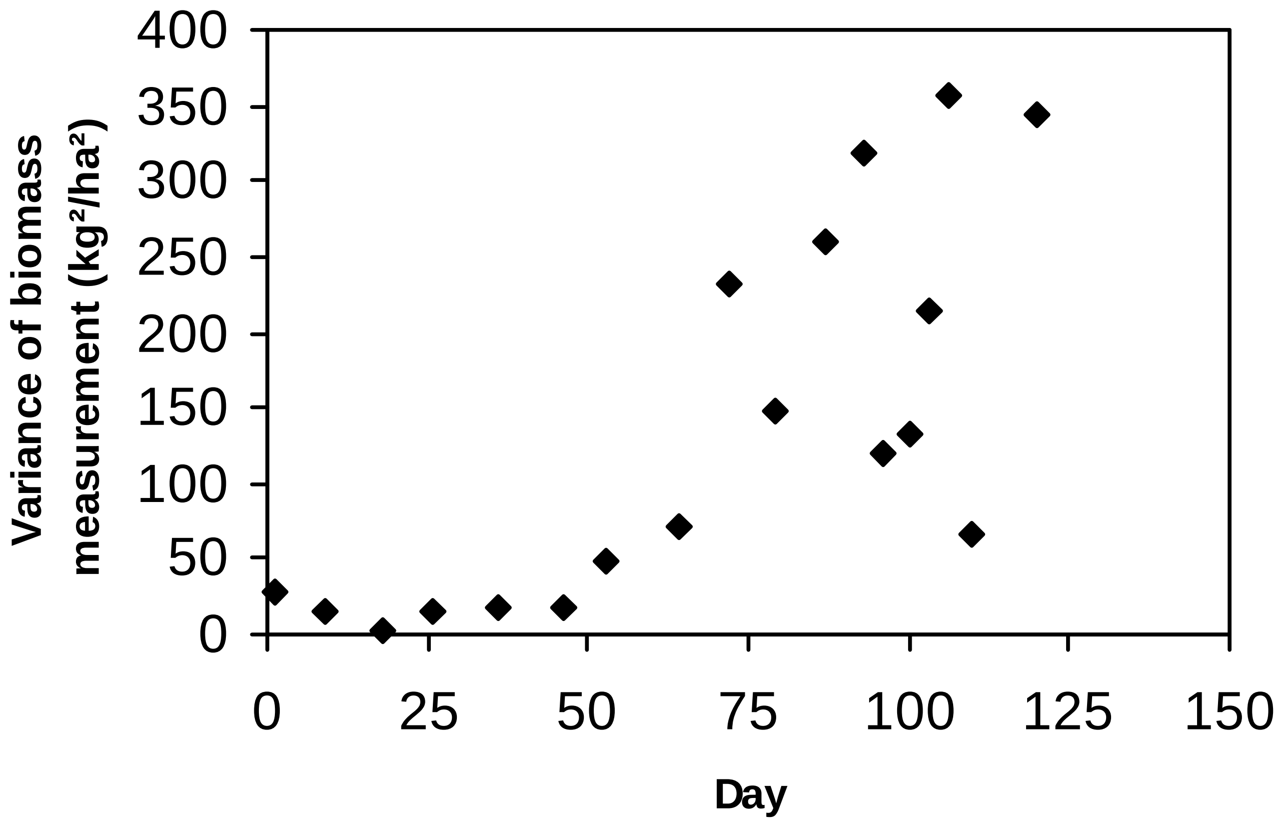
1st option: Ordinary least squares

Find θ minimizing: $$Z(\theta) = \sum_{i=1}^{N} \left[y_i - f\left(t_i, x_i; \theta\right) \right]^2$$

This method performs well when the model error variances are constant

This is not very realistic here.

Here, the variances of the observations are variable



Step 3. Which method?

2nd option: Weighted least squares

Find the value of θ minimizing:

$$Z(\theta) = \sum_{i=1}^{N} \frac{\left[y_i - f(t_i, x_i; \theta) \right]^2}{\sigma_i^2}$$

with $$\hat{\sigma}_i^2 = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(y_{ik} - y_i \right)^2$$

Weighted least squares

minimizing

$$Z_{MCP}(\theta)=\sum_{i=1}^{6}\sum_{j=1}^{10}\frac{\left[y_{ij}^{MS}-f^{MS}\left(t_j,x_i;\theta\right)\right]^2}{\hat{\sigma}_{MS.ij}^2}+\sum_{i=1}^{6}\sum_{j=1}^{10}\frac{\left[y_{ij}^{N}-f^{N}\left(t_j,x_i;\theta\right)\right]^2}{\hat{\sigma}_{N.ij}^2}+\sum_{i=1}^{6}\sum_{j=1}^{10}\frac{\left[y_{ij}^{L}-f^{L}\left(t_j,x_i;\theta\right)\right]^2}{\hat{\sigma}_{L.ij}^2}$$

$$\hat{\sigma}_{MS.ij}^2=\frac{1}{K\left(K-1\right)}\sum_{k=1}^{K}\left[y_{ijk}^{MS}-y_{ij}^{MS}\right]^2$$

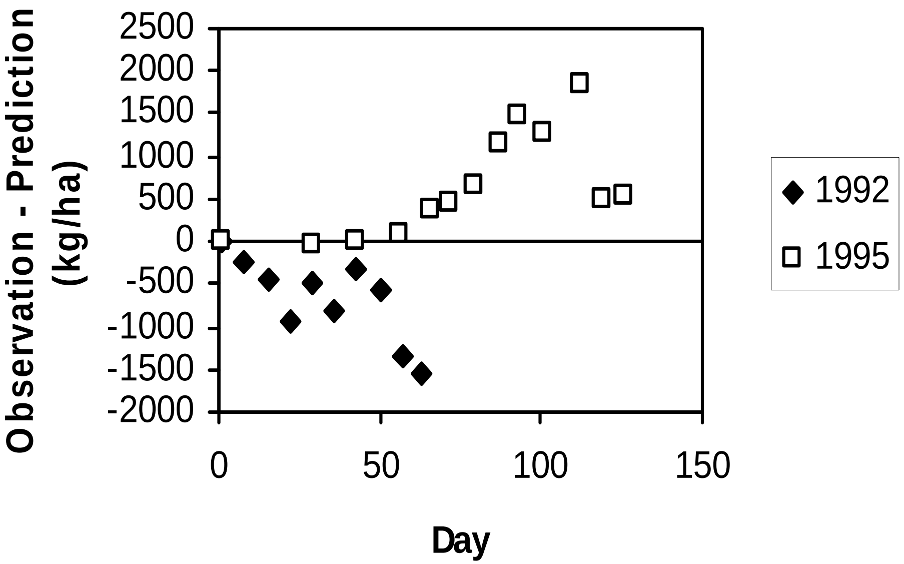
Results of step 3: weighted least square estimates of the 4 model parameters

Parameter	Initial value	Estimated value
E_{BMAX} (g/MJ)	3.3	3.29 (0.11)
D	0.028	0.037 (0.06)
K	0.72	0.74 (0.001)
V_{MAX} (kg/ha/dj)	0.5	0.38 (0.02)

Step 4. Are these estimators accurate?

Model residuals

Residuals are not independant →



Methods for taking correlations into account

- Generalized least squares

- Mixed models

Four estimation problems

Pb.A: One parameter

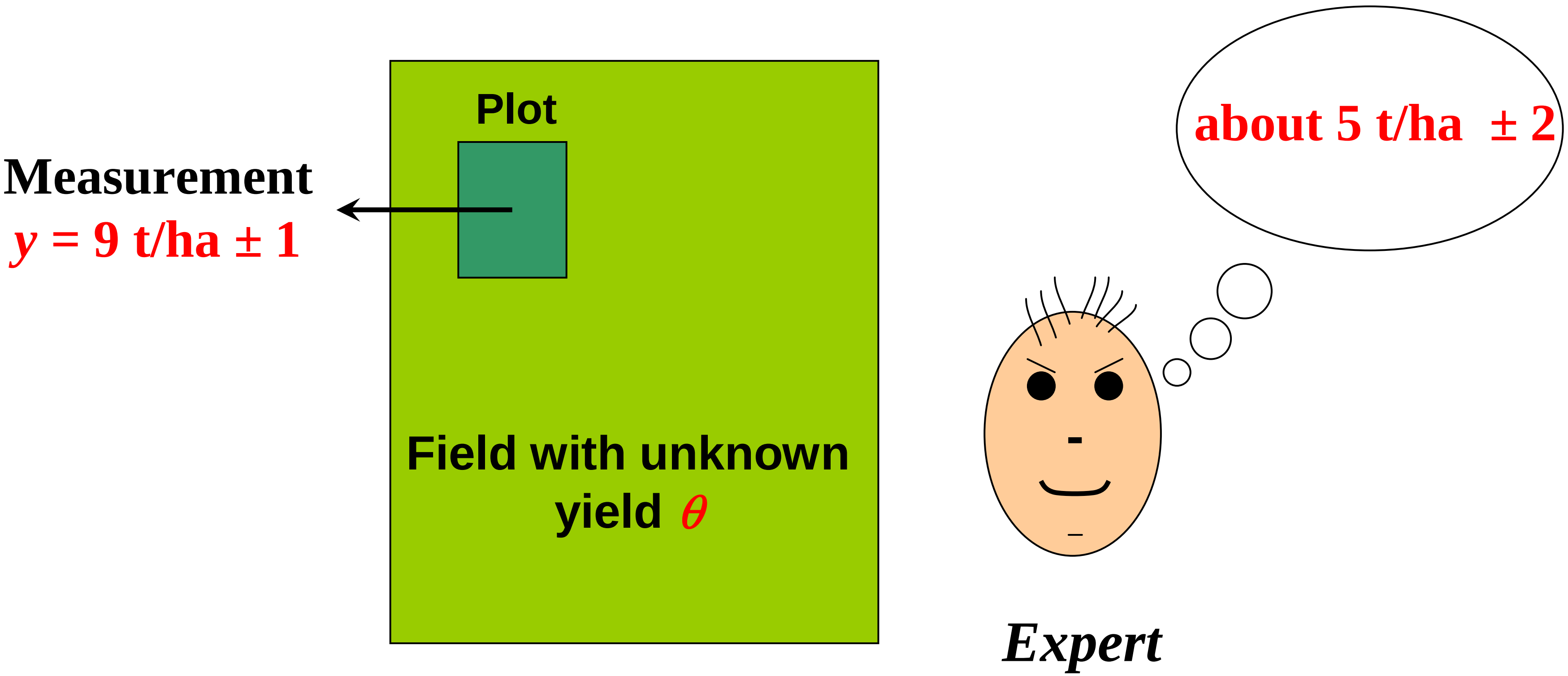
Pb.B: Linear with 2 parameters

Pb.C: Non linear with 18 parameters

Pb.D: Estimation from data and prior information

Problem D

Estimation of crop yield θ by combining a measurement with expert knowledge



Bayesian method

The Bayesian estimation methods allow one to combine information from different sources to estimate unknown parameters

Basic principles:

- Both data and external information (prior) are used
- Computations are based on the Bayes theorem
- Parameters are defined as random variables

Bayes' theorem for model parameters

θ: vector of parameters.

y: vector of observations

P(θ): prior distribution of the parameter values.

P(y | θ): likelihood function.

P(θ | y): posterior distribution of the parameter values.

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{\int_\theta P(y|\theta)P(\theta)}$$

often difficult to compute

How to proceed for estimating the parameters of models

We proceed in three steps:

Step 1: Definition of the ***prior* distribution**.

Step 2: Definition of the **likelihood function**.

Step 3: Computation of the ***posterior* distribution** using Bayes' theorem.

Three distributions in Bayes' theorem

- Prior parameter distribution = probability distribution describing our initial knowledge about parameter values

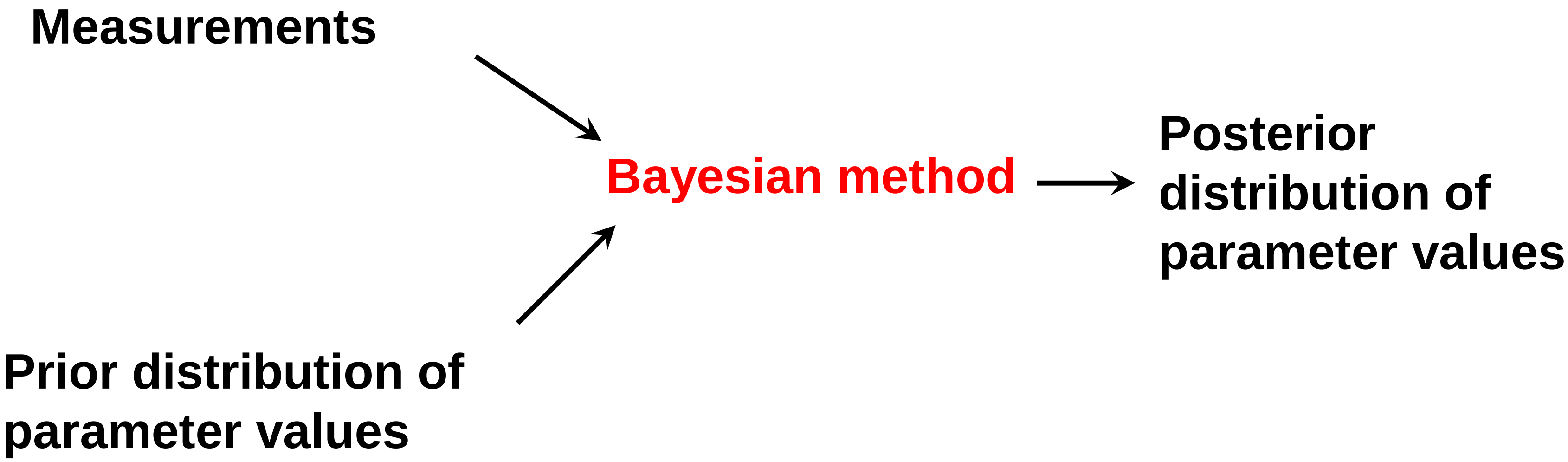
$$P(\theta)$$

- Likelihood function = function relating data to parameters

$$P(y|\theta)$$

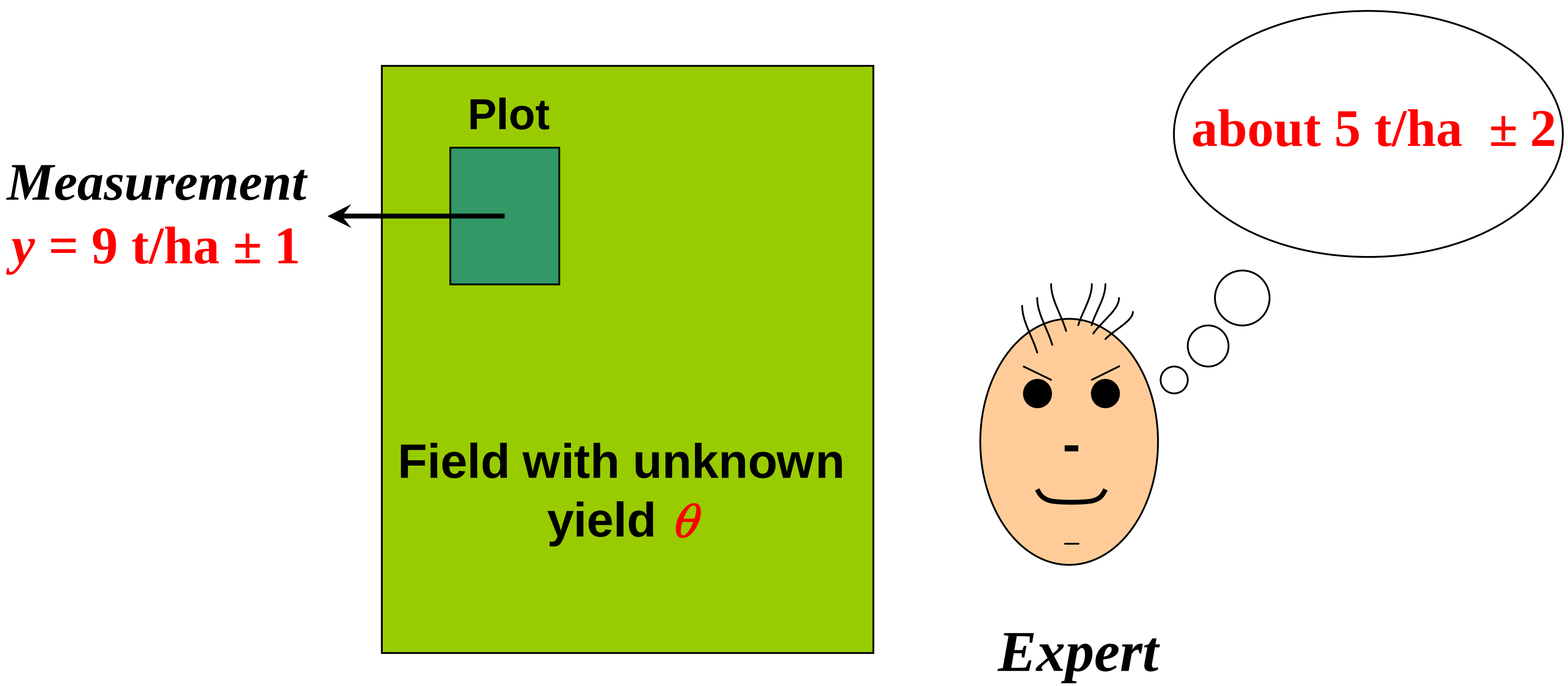
- Posterior parameter distribution = probability distribution summarizing our final state of knowledge about parameters

$$P(\theta|y)$$


Measurements
Bayesian method
Posterior distribution of parameter values
Prior distribution of parameter values

Our problem

Estimation of crop yield θ by combining a measurement with expert knowledge



Prior distribution

- It describes our belief about the parameter values **before** we observe the measurements.

- It is based on past studies, expert knowledge, and litterature.

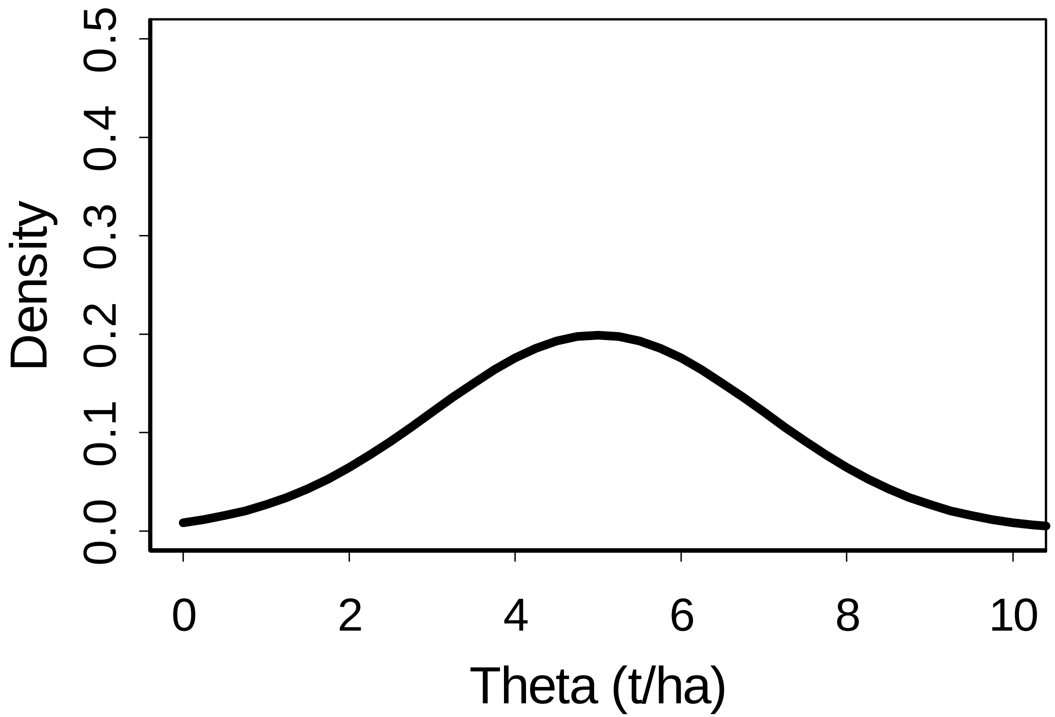
Definition of a prior distribution for our problem

$$\theta \sim N(\mu, \tau^2)$$

$$P(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\tau^2}\right] = \frac{1}{\sqrt{2\pi 4}} \exp\left[-\frac{(\theta-5)^2}{2\times 4}\right]$$

- Normal probability distribution.
- Expected value equal to 5 t/ha.
- Standard error equal to 2 t/ha

Plot of the prior distribution



Statistical model and likelihood function

$$y = \theta + \varepsilon \qquad \text{with } \varepsilon \sim N(0, \sigma^2)$$

$$y \mid \theta \sim N(\theta, \sigma^2)$$

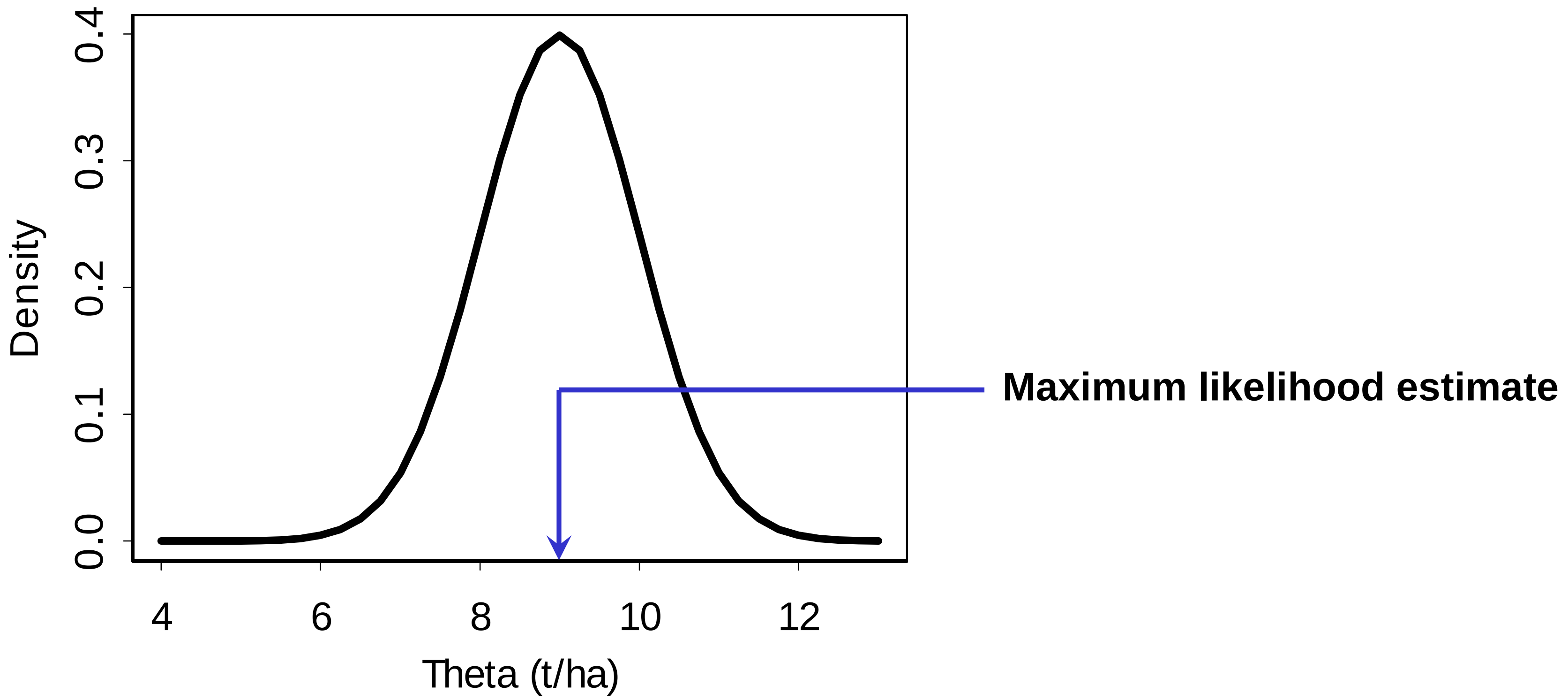
Definition of a likelihood function

$$y \mid \theta \sim N(\theta, \sigma^2)$$

$$P(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\theta)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(9-\theta)^2}{2}\right]$$

- Normal probability distribution.
- Measurement *y* assumed unbiaised and equal to 9 t/ha.
- Standard error σ assumed equal to 1 t/ha

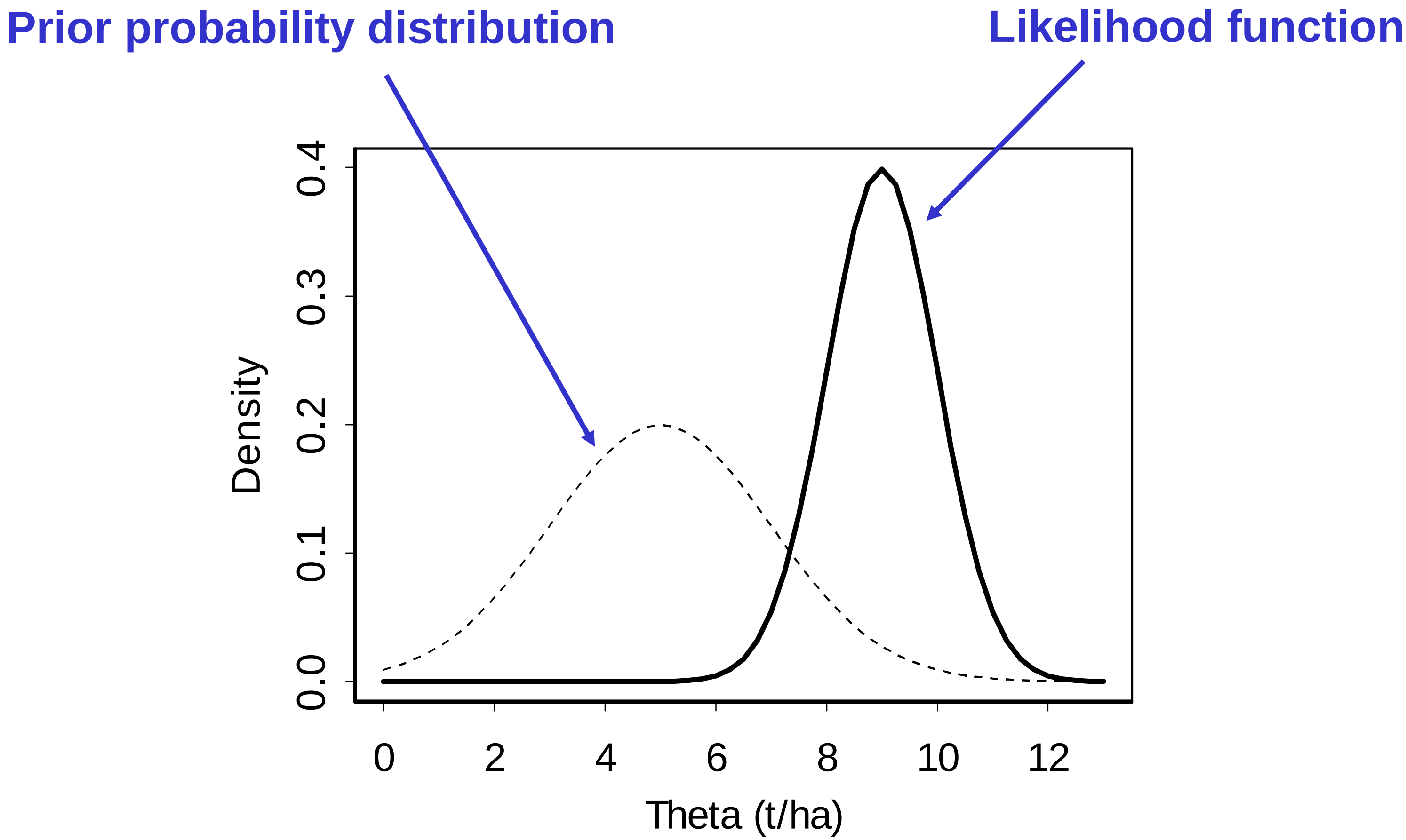
Definition of a likelihood function



Maximum likelihood

Likelihood functions are also used by frequentist to implement the *maximum likelihood method*.

The maximum likelihood estimator is the value of θ maximizing $P(y \mid \theta)$.


Prior probability distribution
Likelihood function
Density
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
12
Theta (t/ha)

Example (continued)

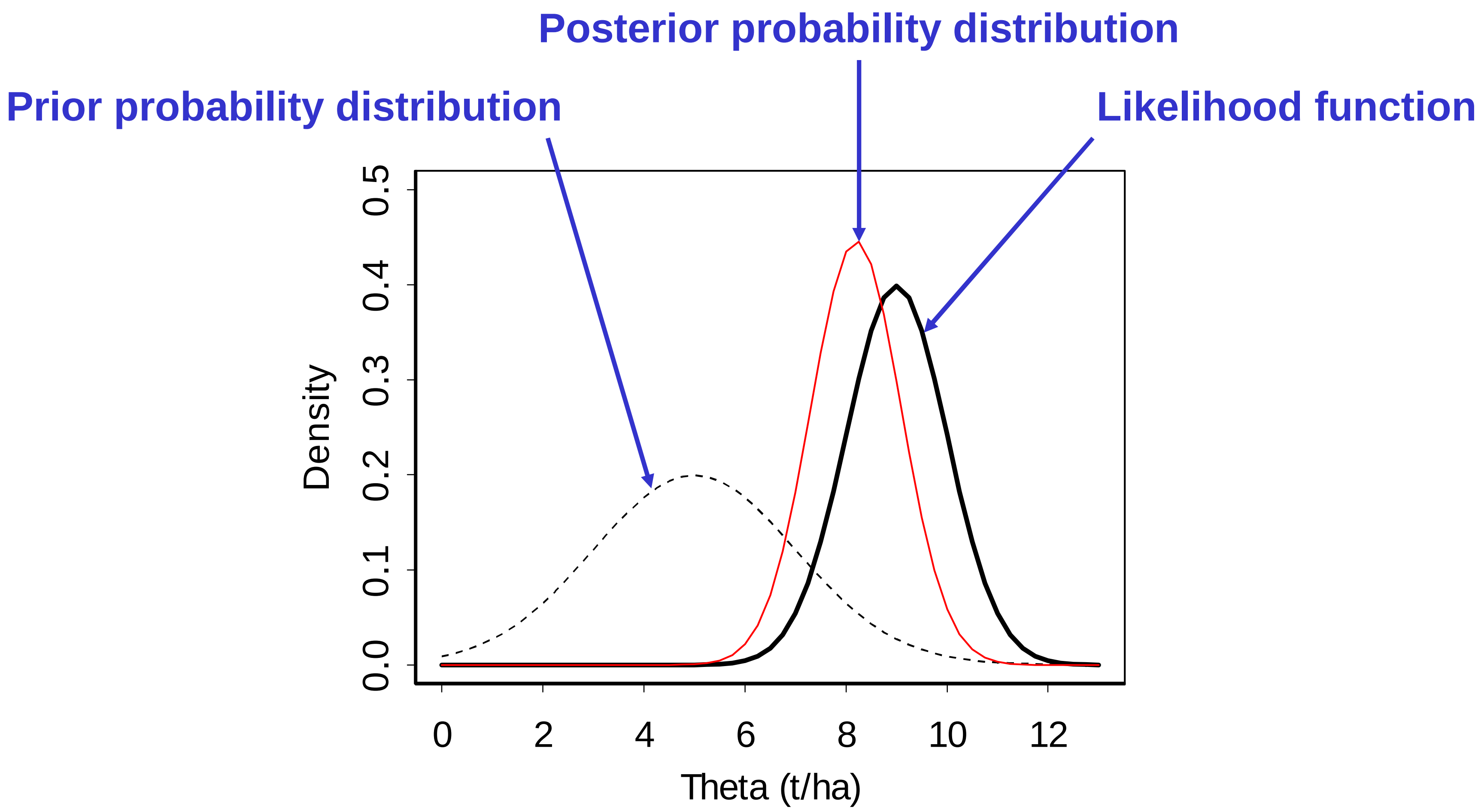
Analytical expression of the posterior distribution

$$\theta \mid y \sim N(\mu_{post},\ \tau_{post}^2)$$

$$\mu_{post} = (1-w)\times\mu + w\times y = 8.2$$

$$\tau_{post}^2 = (1-w)\times\tau^2 = 0.8$$

$$w = \frac{\tau^2}{\tau^2+\sigma^2} = \frac{4}{5}$$



Posterior probability distribution
Prior probability distribution
Likelihood function
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
12
Theta (t/ha)

Discussion of the posterior distribution

1. Result is a probability distribution (posterior distr.)
2. Posterior mean is intermediate between prior mean and observation.
3. Weights depend on prior variance and measurement error.
4. Posterior variance is lower than both prior variance and measurement error variance.
5. A full distribution was derived from only one data

Frequentist *versus* Bayesian

Bayesian analysis introduces an element of subjectivity:
***the prior distribution*.**

But its representation of the uncertainty is easy to understand

- the uncertainty is assessed conditionally to the observations,

- the calculations are straightforward when the posterior distribution is known.

Practical considerations

• The analytical expression of the posterior distribution can be derived for simple applications.

• For complex problems, the posterior distribution must be **approximated**

→ Markov chain Monte Carlo algorithms (MCMC)

→ Importance sampling

• Softwares were recently developed for running MCMC (WinBUGS, JAGS…)

Conclusion. Proceed in four steps

1. How many and which parameters should be estimated?

→ In simple models, all parameters.

→ In complex models, parameters must be selected.

2. What kind of information is available?

→ Data

→ Prior information (expert knowledge, literature etc.)

3. Which estimation method?

→ Ordinary least squares,

→ Weighted/Generalized least squares,

→ Bayesian method

4. What is the accuracy of the parameter estimator?

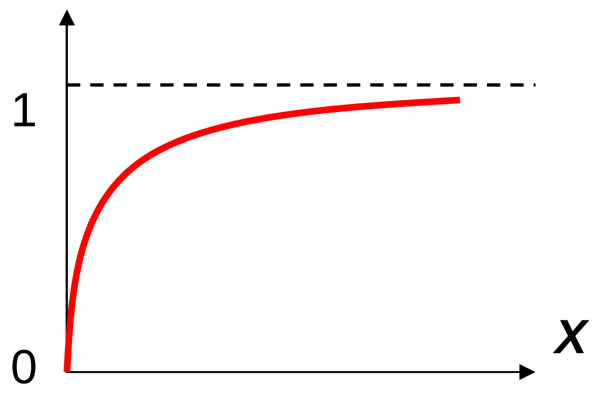
→ Theoretical consideration, variances, residuals.

Exercise

Estimation of the parameter of a non linear model using a Bayesian method

- Non linear model predicting relative yield as a function of a factor *x* (amount of soil mineral N)

$$f(x,\theta)=\left[1-\exp(-\theta\times x)\right]$$



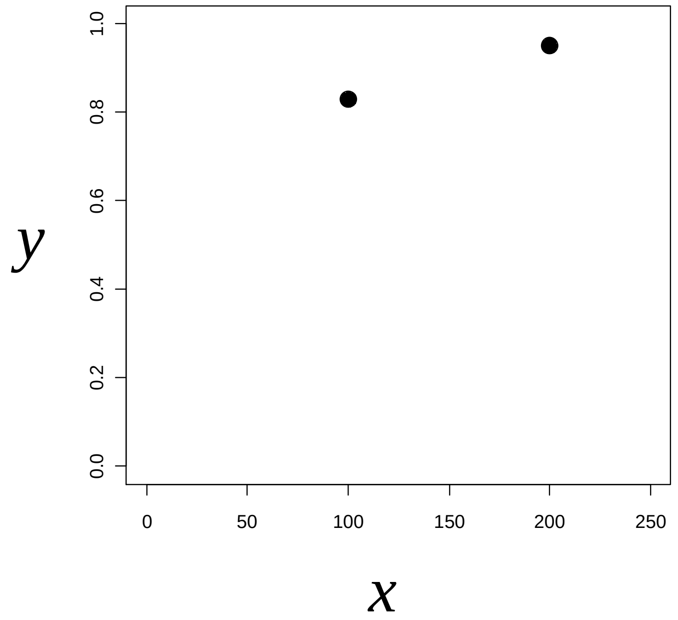
- One parameter θ: the growth rate

Data

Two measurements of relative yield y_1 and y_2 are available:

$y_1 = 0.83$ for $x_1 = 100$ kg/ha

$y_2 = 0.95$ for $x_2 = 200$ kg/ha



Questions

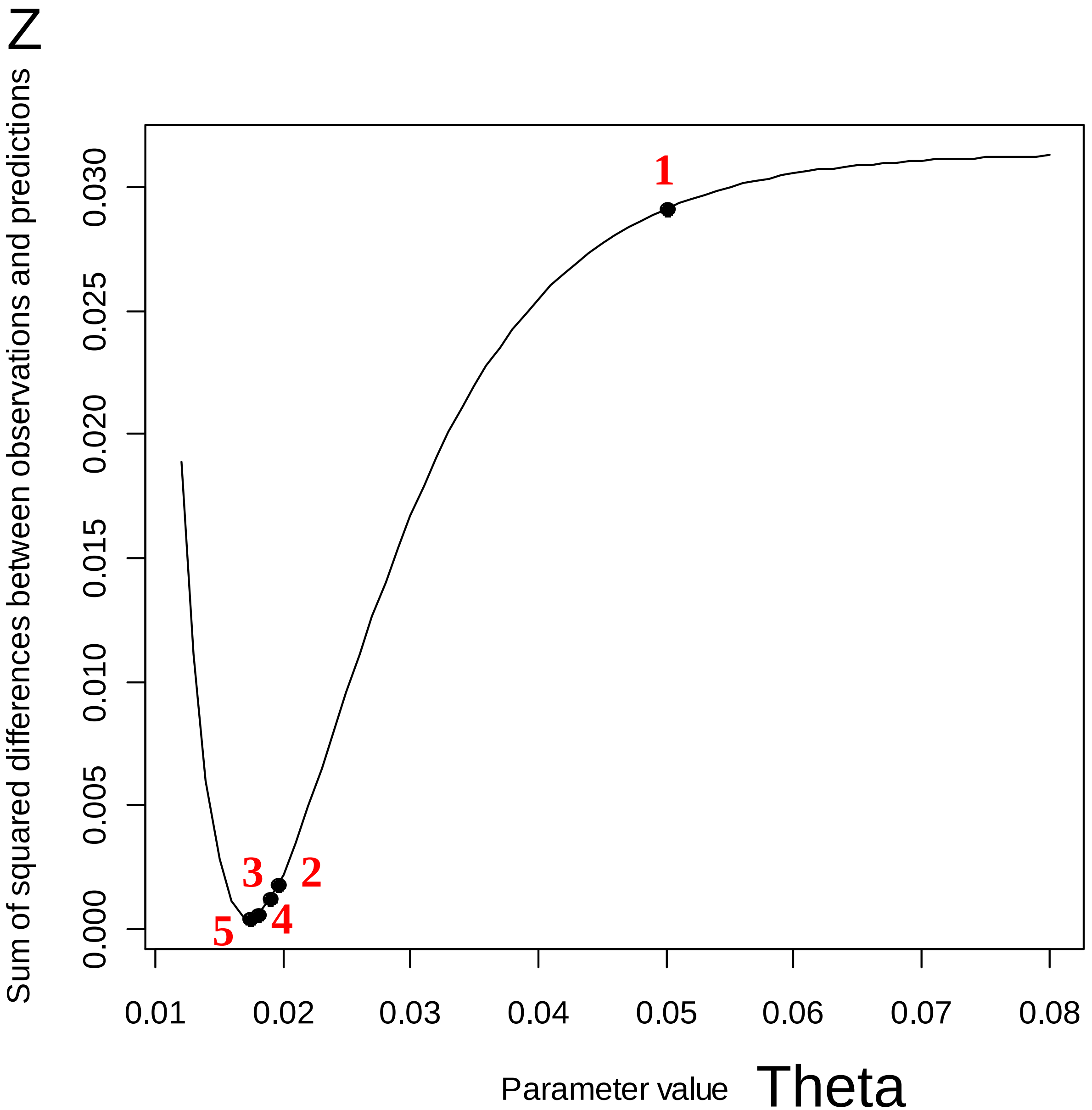
- Estimate the parameter by ordinary least squares
- Estimate the parameter by using a Bayesian method

Ordinary least squares

```
x<-c(100, 200)

y<-c(0.83, 0.95)

TAB<-data.frame(x,y)
```

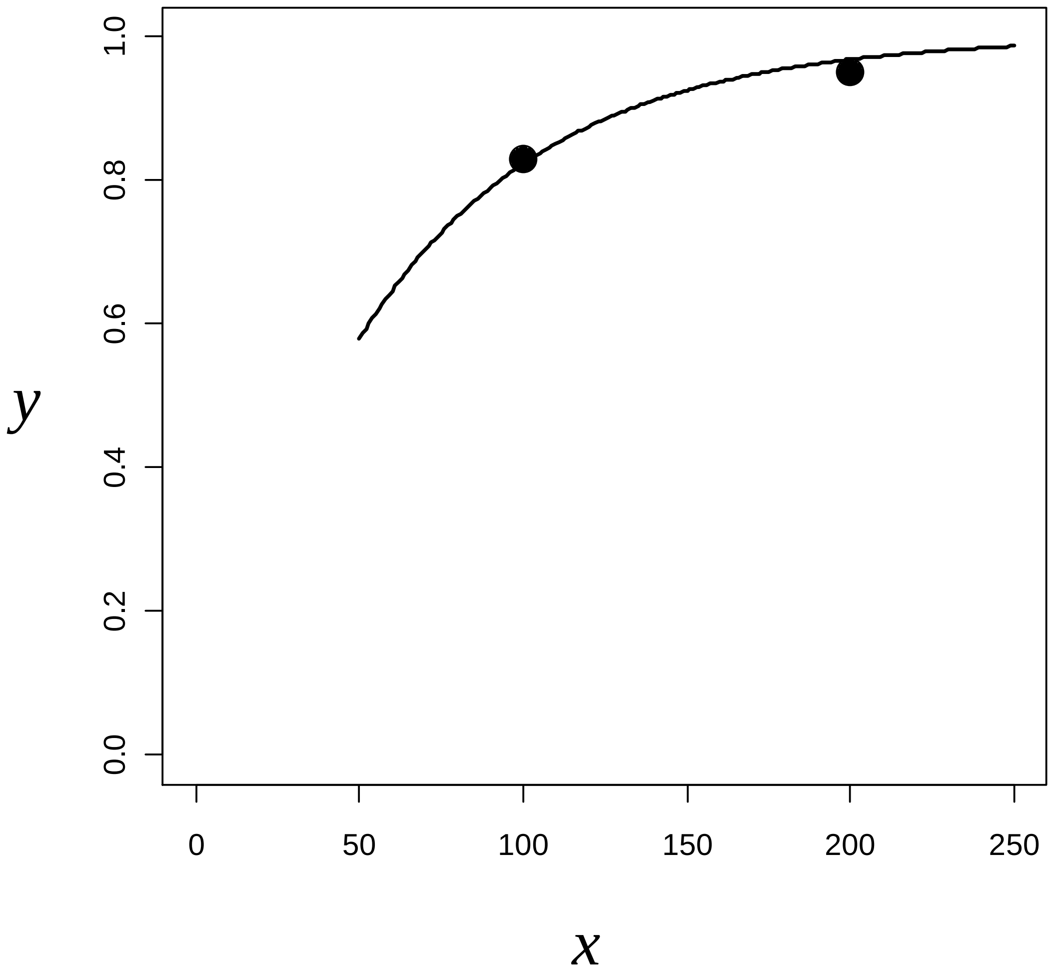
```
x<-c(100, 200)

y<-c(0.83, 0.95)

TAB<-data.frame(x,y)

Fit<-nls(y~1-exp(-Theta*x), data=TAB, start=list(Theta=0.05), trace=T)

print(summary(Fit))
```

```
> Fit<-nls(y~1-exp(-Theta*x), data=TAB, start=list(Theta=0.05), trace=T)
0.02914996 :  0.05
0.001751208 :  0.01959284
0.001160448 :  0.01897615
0.0005836208 :  0.01810939
0.0003974804 :  0.01732987
0.0003974661 :  0.01733614
0.0003974661 :  0.01733635

> print(summary(Fit))

Formula: y ~ 1 - exp(-Theta * x)

Parameters:
      Estimate Std. Error t value Pr(>|t|)
Theta 0.017336   0.001064   16.29    0.039 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01994 on 1 degrees of freedom

Number of iterations to convergence: 6
```


Z
Sum of squared differences between observations and predictions
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Parameter value
Theta
1
2
3
4
5

```
X.vec<-50:250

Y.vec<-1-exp(-coef(Fit)[1]*X.vec)

plot(x,y, xlim=c(0, 250), pch=19, cex=2, ylim=c(0,1))

lines(X.vec, Y.vec, lwd=2)
```

$$\hat{\theta} = 0.0173$$



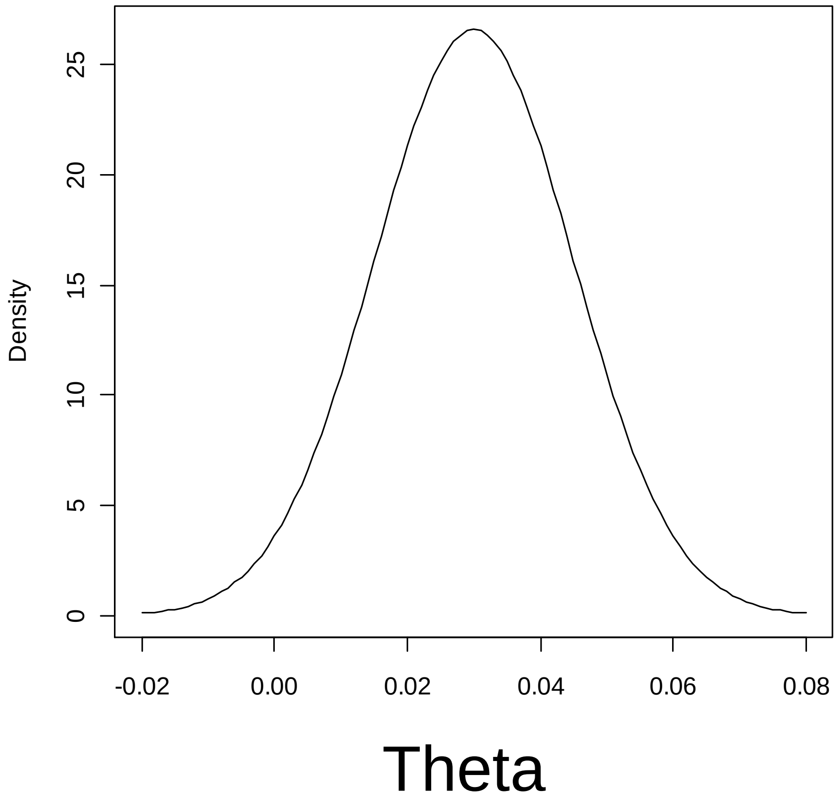
A Bayesian approach

Prior distribution

Assume that the prior distribution is defined from expert knowledge as:

$\theta \sim N(\mu, \tau^2)$

$\theta \sim N(0.03, 0.015^2)$



Data

- Two measurements of relative yield y_1 and y_2 are available:

  $y_1 = 0.83$ for $x_1 = 100$ kg/ha,

  $y_2 = 0.95$ for $x_2 = 200$ kg/ha.

Statistical model

The statistical model is defined as

$y = f(x, \theta) + \varepsilon$

$\varepsilon \sim N(0, \sigma^2)$

with independance between the two values of ε

Likelihood

$$P\left(y_1, y_2 \middle| \theta\right) = P\left(y_1 \middle| \theta\right) \times P\left(y_2 \middle| \theta\right)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\left[y_1 - f\left(x, \theta\right)\right]^2}{2\sigma^2}\right\} \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\left[y_2 - f\left(x, \theta\right)\right]^2}{2\sigma^2}\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\left[y_1 - \left(1 - \exp\left(-\theta \times x_1\right)\right)\right]^2}{2\sigma^2}\right\} \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\left[y_2 - \left(1 - \exp\left(-\theta \times x_2\right)\right)\right]^2}{2\sigma^2}\right\}$$

We assume that σ is known and equal to 0.02.

Importance sampling

Step 0: **Use the prior distribution as the proposal distribution.**

$g(\theta) = N(0.03, 0.015^2)$.

Step 1: **Generate N parameter values $\theta_1, \theta_2, \ldots, \theta_N$.**

Step 2: **Calculate a « weight » for each parameter value $w_1, w_2, \ldots, w_N$.**

The weight is equal to the likelihood value, $w_i = P(y_1 | \theta_i) \times P(y_2 | \theta_i)$

$$w_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left[Y_1 - \left(1 - \exp\left(-\theta_i \times X_1\right)\right)\right]^2}{2\sigma^2} \right\} \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left[Y_2 - \left(1 - \exp\left(-\theta_i \times X_2\right)\right)\right]^2}{2\sigma^2} \right\}$$

Step 3: **Calculate normalized weights $w_1^*, \ldots, w_N^*$.**

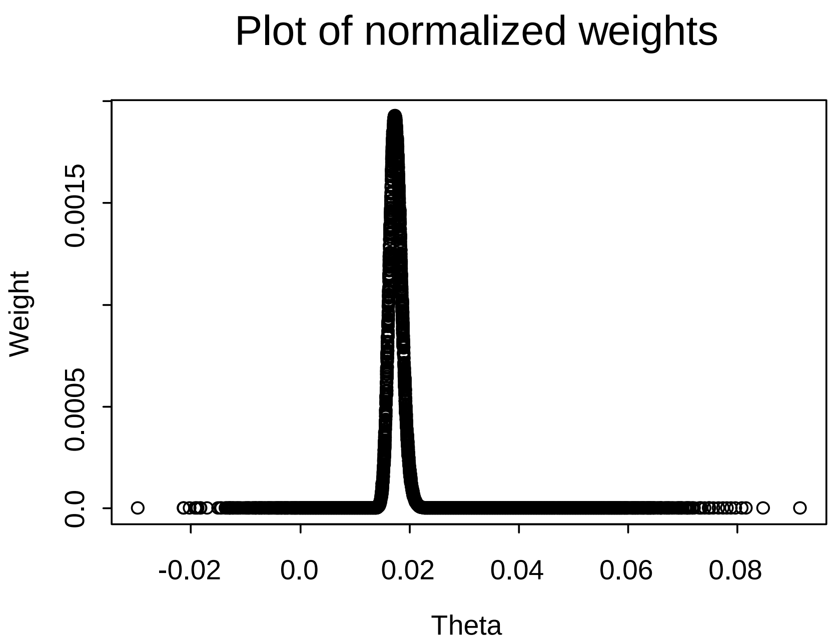
$$w_i^* = \frac{w_i}{\sum_{i=1}^{N} w_i}$$

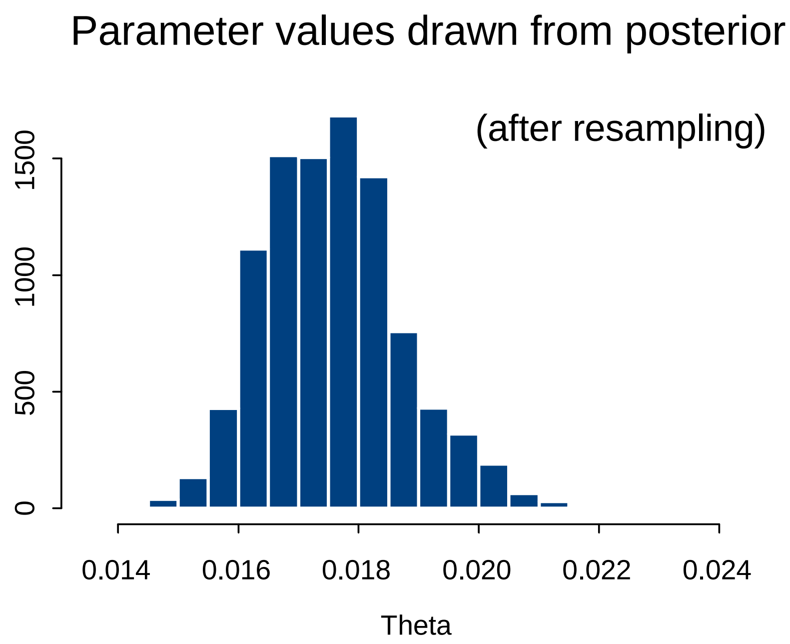
Step 4: **Use the sample of parameter values and the normalized weights to approximate the posterior distribution.**

Implementation

- The algorithm is implemented with the R function ISgrowth.txt.
- You can run it yourself and modify it for other models

Results obtained with *N*=10000





Estimated posterior mean: 0.0176

Estimated posterior standard deviation: 1.17 10^{-3}

Which value for *N*?

- The algorithm is run five times (with different seeds) for two different *N* values:

  N=100

  N=10000

- The posterior mean and posterior variance are computed after each run.

- The stability of the result is analyzed.

How many simulations ?

N	Run	Posterior mean	Posterior standard deviation
100	1	0.0178	9.99E-04
100	2	0.0173	9.48E-04
100	3	0.0173	1.10E-03
100	4	0.0176	1.04E-03
100	5	0.0176	9.91E-04
10000	1	0.0176	1.17E-03
10000	2	0.0176	1.14E-03
10000	3	0.0176	1.17E-03
10000	4	0.0176	1.14E-03
10000	5	0.0176	1.14E-03

The estimation of the posterior mean is very accurate with *N*=10000.