Estimation of model parameters

David Makowski
INRA
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, literature)

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

**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 $\theta$. 
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 »
Step 4. Is the estimator accurate?

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

Mean squared error \quad \text{Bias}^2 \quad \text{Variance}
Step 4. Is the estimator accurate?

a. Theoretical consideration

« Under some assumptions, our estimator is *unbiased* and of *minimum variance* among the unbiased estimators »
Step 4. Is the estimator accurate?

b. Estimator variance

\[ \text{var}(\hat{\theta}) \text{ can be estimated from data} \]

Example:

- If \( y_1 = 30, y_2 = 39 \) and \( y_3 = 35 \), the estimated variance is \( 6.78 \text{ q}^2/\text{ha}^2 \), standard deviation=\( 2.6 \text{ q/ha} \).

- If \( y_1 = 32, y_2 = 38 \) and \( y_3 = 39 \), the estimated variance is \( 4.78 \text{ q}^2/\text{ha}^2 \), standard deviation=\( 2.19 \text{ q/ha} \).
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: $\theta_1$ and $\theta_2$
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 $\theta_1$ and $\theta_2$ minimizing

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

that is

$$\hat{\theta}_2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})(x_i - \bar{X})}{\sum_{i=1}^{N} (x_i - \bar{X})^2}$$

$$\hat{\theta}_1 = \bar{Y} - \hat{\theta}_2 \bar{X}.$$
$y_5 - (\theta_1 + \theta_2 x_5)$
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?

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

- Mean squared error
- Bias$^2$
- 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 $\text{var}(\hat{\theta})$ from the data

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

$$\sqrt{\text{var}(\hat{\theta}_2)} = 0.09 \text{ kg.kg}^{-1}$$
Step 4. Are these estimators accurate?

c. Analysis of the residuals

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

Useful to check the independance of the model errors and variance homogeneity
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: \( \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p \)

  \( \Rightarrow \) Analytic relationships between estimators and data

• Number of data > Number of parameters

• Only one type of measurements

• No prior information about parameter values

• Softwares are available for the computations (SAS, R, ModelMaker...).
It can be much more difficult

• Non linear models: \( \neq \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p \)
  \( \rightarrow \) No analytic relationship between the estimators and the data

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

• Complex dataset

  \( \rightarrow \) 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) \( \rightarrow MS_t \)
- Nitrogen uptake (kg/ha), \( \rightarrow QN_t \)
- Leaf Area Index \( \rightarrow LAI_t \)

Input variables:
- Global daily radiation \( \rightarrow RG_t \)
- Average daily temperature, \( \rightarrow T_t \)
- Initial values of biomass and nitrogen uptake \( \rightarrow MS_0, QN_0 \)
Simulations of wheat biomass using the AZODYN dynamic crop model
Problem C

Few model equations

\[ MS_j = MS_{j-1} + \left( E_{b_{\text{max}}} \times ft_{j-1} \times Ei_{j-1} \times C \times RG_{j-1} \right) \]

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

\[ LAI_{j-1} = D \times QNc_{j-1} \]

\[ MS_j = MS_{j-1} + \left\{ E_{b_{\text{max}}} \times C \times E_{i_{\text{max}}} \left[ 1 - \exp\left(-K \times D \times QNc_{j-1}\right)\right] \times ft_{j-1} \times RG_{j-1} \right\} \]
### 18 parameters

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

1: Dynamic system model

\[ MS_t = MS_{t-1} + g(X_{t-1}; \theta) \]

2: Response model

\[ MS_t = f(t, X; \theta) \]
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»
\[ MS_j = MS_{j-1} + \left( E_{b\text{max}} \times ft_{j-1} \times Ei_{j-1} \times C \times RG_{j-1} \right) \]

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

\[ LAI_{j-1} = D \times QNc_{j-1} \]

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

Case 1: only observed values of MS are available

Case 2: observed values of MS and LAI are available
\[
MS_j = MS_{j-1} + \left( E_{b_{\text{max}}} \times ft_{j-1} \times E_{i_{j-1}} \times C \times RG_{j-1} \right)
\]

\[
E_{i_{j-1}} = E_{i_{\text{max}}} \left[ 1 - \exp \left( -K \times LAI_{j-1} \right) \right]
\]

\[
LAI_{j-1} = D \times QNc_{j-1}
\]

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

**Case 1: only observed values of MS are available**

- the 3 parameters \( E_{b_{\text{max}}}, C, E_{i_{\text{max}}} \)

It is not possible to estimate simultaneously

- 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 \( E_{b_{\text{max}}}, C, E_{i_{\text{max}}} \)
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 »

<table>
<thead>
<tr>
<th>Number of estimated parameters</th>
<th>MSEPcv</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MSEP₁</td>
</tr>
<tr>
<td>2</td>
<td>MSEP₂</td>
</tr>
<tr>
<td>3</td>
<td>MSEP₃</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>P</td>
<td>MSEPₚ</td>
</tr>
</tbody>
</table>
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: $E_{\text{BMAX}}$, $D$, $K$ and $V_{\text{MAX}}$
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 $\theta$ 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 $\theta$ minimizing:

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

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 $\theta$ minimizing:

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

with

$$\hat{\sigma}_i^2 = \frac{1}{K(K-1)} \sum_{k=1}^{K} (y_{ik} - y_i)^2$$
Weighted least squares

minimizing

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

\[
\hat{\sigma}^2_{MS,ij} = \frac{1}{K (K - 1)} \sum_{k=1}^{K} \left[ y_{ijk}^{MS} - y_{ij}^{MS} \right]^2
\]
Results of step 3: weighted least square estimates of the 4 model parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Initial value</th>
<th>Estimated value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{BMAX}$ (g/MJ)</td>
<td>3.3</td>
<td>3.29 (0.11)</td>
</tr>
<tr>
<td>D</td>
<td>0.028</td>
<td>0.037 (0.06)</td>
</tr>
<tr>
<td>K</td>
<td>0.72</td>
<td>0.74 (0.001)</td>
</tr>
<tr>
<td>$V_{MAX}$ (kg/ha/dj)</td>
<td>0.5</td>
<td>0.38 (0.02)</td>
</tr>
</tbody>
</table>
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 $\theta$ by combining a measurement with expert knowledge

- Measurement
  $y = 9 \text{ t/ha} \pm 1$

- Field with unknown yield $\theta$

- Expert
  about $5 \text{ t/ha} \pm 2$
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

\( \theta \): vector of parameters.

\( y \): vector of observations

\( P(\theta) \): prior distribution of the parameter values.

\( P(y | \theta) \): likelihood function.

\( P(\theta | y) \): 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

Prior distribution of parameter values

Bayesian method

Posterior distribution of parameter values
Our problem

Estimation of crop yield $\theta$ by combining a measurement with expert knowledge

Measurement

$y = 9 \text{ t/ha} \pm 1$

Field with unknown yield $\theta$

about $5 \text{ t/ha} \pm 2$

Expert
Prior distribution

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

• It is based on past studies, expert knowledge, and literature.
Definition of a prior distribution for our problem

\[ \theta \sim N(\mu, \tau^2) \]

\[ P(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[ -\frac{(\theta - \mu)^2}{2\tau^2} \right] = \frac{1}{\sqrt{2\pi}4} \exp\left[ -\frac{(\theta - 5)^2}{2\times4} \right] \]

- 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

\[ y = \theta + \varepsilon \quad \text{with} \quad \varepsilon \sim N(0, \sigma^2) \]

\[ y \mid \theta \sim N(\theta, \sigma^2) \]
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 unbiased and equal to 9 t/ha.
• Standard error \( \sigma \) 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 $\theta$ maximizing $P(y \mid \theta)$.
Prior probability distribution

Likelihood function

Theta (t/ha)

Density
Example (continued)

Analytical expression of the posterior distribution

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

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

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

\[ w = \frac{\tau^2}{\tau^2 + \sigma^2} = \frac{4}{5} \]
Prior probability distribution

Likelihood function

Posterior probability distribution
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 $\theta$: the growth rate
Data

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

$$y_1 = 0.83 \text{ for } x_1 = 100 \text{ kg/ha}$$

$$y_2 = 0.95 \text{ for } x_2 = 200 \text{ kg/ha}$$

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
Parameter value

Sum of squared differences between observations and predictions

Theta
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
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:

  \[ y_1 = 0.83 \text{ for } x_1 = 100 \text{ kg/ha}, \]
  \[ y_2 = 0.95 \text{ for } x_2 = 200 \text{ kg/ha}. \]
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 $\varepsilon$
Likelihood

\[ P(y_1, y_2 | \theta) = P(y_1 | \theta) \times P(y_2 | \theta) \]

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

We assume that \( \sigma \) is known and equal to 0.02.
Importance sampling

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

\[ g(\theta) = \text{N}(0.03, 0.015^2). \]

**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, \( w_i = P(y_1|\theta_i) \times P(y_2|\theta_i) \)

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

**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 $N=10000$

Plot of normalized weights

Parameter values drawn from posterior (after resampling)

Estimated posterior mean: 0.0176
Estimated posterior standard deviation: $1.17 \times 10^{-3}$
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?

<table>
<thead>
<tr>
<th>N</th>
<th>Run</th>
<th>Posterior mean</th>
<th>Posterior standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1</td>
<td>0.0178</td>
<td>9.99E-04</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.0173</td>
<td>9.48E-04</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.0173</td>
<td>1.10E-03</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.0176</td>
<td>1.04E-03</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.0176</td>
<td>9.91E-04</td>
</tr>
<tr>
<td>10000</td>
<td>1</td>
<td>0.0176</td>
<td>1.17E-03</td>
</tr>
<tr>
<td>10000</td>
<td>2</td>
<td>0.0176</td>
<td>1.14E-03</td>
</tr>
<tr>
<td>10000</td>
<td>3</td>
<td>0.0176</td>
<td>1.17E-03</td>
</tr>
<tr>
<td>10000</td>
<td>4</td>
<td>0.0176</td>
<td>1.14E-03</td>
</tr>
<tr>
<td>10000</td>
<td>5</td>
<td>0.0176</td>
<td>1.14E-03</td>
</tr>
</tbody>
</table>

The estimation of the posterior mean is very accurate with \( N=10000 \).