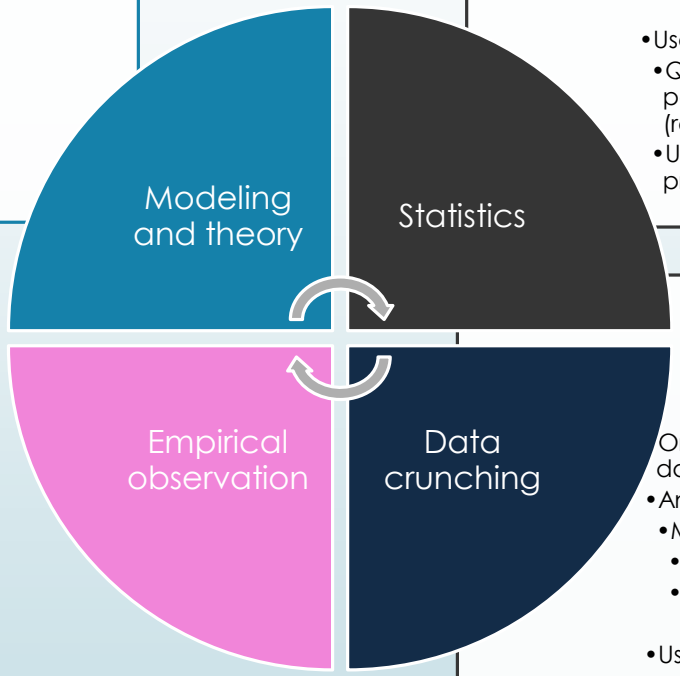
Model fitting
(and Monte-Carlo simulations)

BIOS13 Modelling biological systems
Department of Biology, Lund University
Mikael Pontarp

Outline

- Stochastic modelling in general and stochastic population dynamics in particular (exercise preparation)
- Least squares method
- Non-linear model fitting in R
- Model choice, the AIC
- Running Monte-Carlo simulations



- Phenomenological models
- Mechanistic models
- Process based model

- Used for:
  - Conceptualization
  - Understanding
  - Prediction
  - Develop theory

- Linear models
- Models of variation
- Models of fixed and random effects

- Used for:
  - Quantifying observed patterns and relationships (response vs explanatory)
  - Understanding and predict

Organize, clean and save data

- Analyze data
  - Models
    - Statistics
    - Algorithms

- Used for:
  - Making info in data accessible
  - Quantifying and understanding observed patterns

Stochastic modelling

- The art of stochastic modelling usually boils down to deciding what is stochastic and which distribution does it follow.
- Most often, there are many plausible alternatives.
- Queue example:
  Is both entering and leaving the queue a stochastic process or is it sufficient to make the leaving process stochastic?
  What should be the distribution of times it takes to finish an errand?
- The decisions you make may or may not influence the outcome of the model. This is tricky! Knowing which decisions are crucial, and which are not, requires experience (if at all possible).
- Some sort of robustness check is often useful. Try different assumptions.

The stochastic Ricker equation

- The Ricker equation predicts next year's population size, N_{t+1}, based on current population size, N_t:

$$N_{t+1} = N_t e^{r_0\left(1-\frac{N_t}{K}\right)}$$

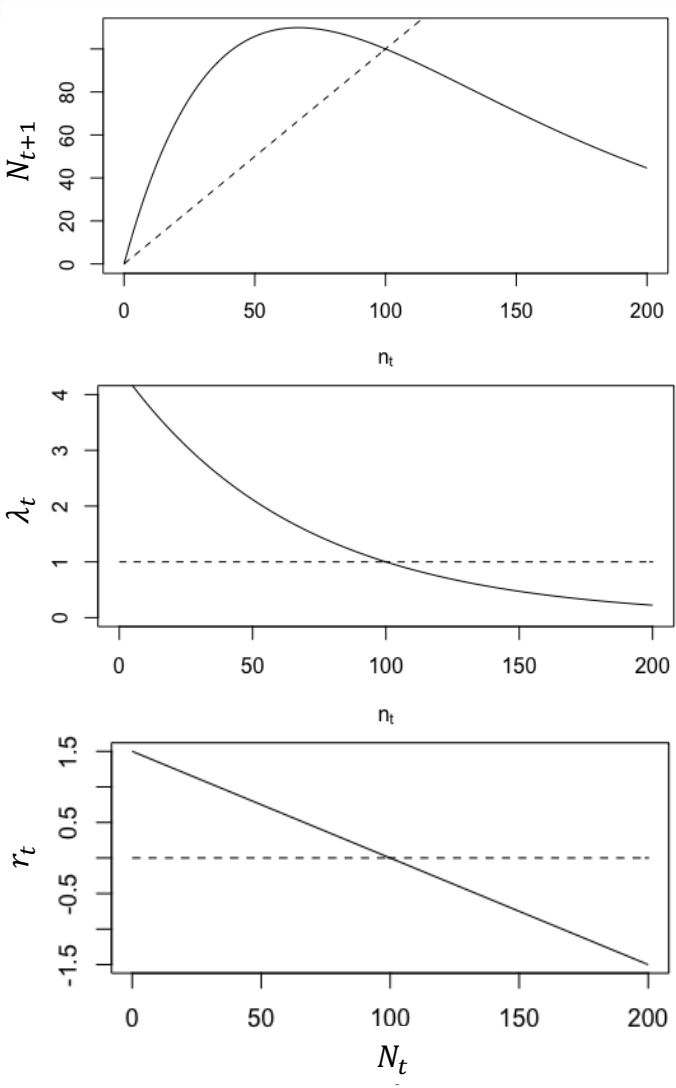
- Put another way, the Ricker equation predicts a *growth rate*, λ_t, defined as the ratio between two consequitive population sizes:

$$\lambda_t = \frac{N_{t+1}}{N_t} = e^{r_0\left(1-\frac{N_t}{K}\right)}$$
$$N_{t+1} = N_t \lambda_t$$

- Put *another* way, the Ricker equation predicts the *exponential growth rate*, r_t, defined as $\ln(\lambda_t)$:

$$r_t = \ln(\lambda_t) = \ln\left(\frac{N_{t+1}}{N_t}\right) = r_0\left(1 - \frac{N_t}{K}\right)$$
$$N_{t+1} = N_t e^{r_t}$$

Note: Any population dynamic model predicts these things, but the functions don't look the same.



What is stochastic?

Based on the previous slide, there are at least three ways to add 'stochastic environmental fluctuations' to the Ricker equation:

1. As a random deviation of population size, directly:

$$N_{t+1} = N_t e^{r_0\left(1-\frac{N_t}{K}\right)} + \varepsilon_t$$

2. As a random deviation of the *growth rate*, λ_t:

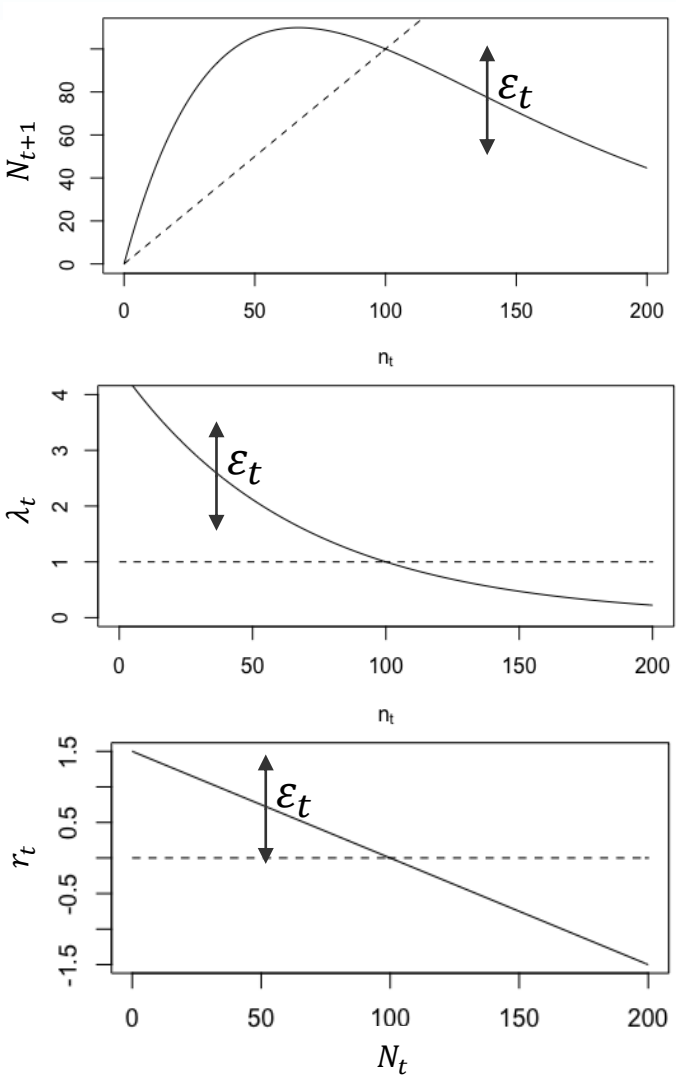
$$\lambda_t = \frac{N_{t+1}}{N_t} = e^{r_0\left(1-\frac{N_t}{K}\right)} + \varepsilon_t$$

$$N_{t+1} = N_t\lambda_t = N_t\left(e^{r_0\left(1-\frac{N_t}{K}\right)} + \varepsilon_t\right)$$

3. As a random deviation of the *exponential growth rate*, r_t:

$$r_t = r_0\left(1-\frac{N_t}{K}\right) + \varepsilon_t$$

$$N_{t+1} = N_t e^{r_t} = N_t e^{r_0\left(1-\frac{N_t}{K}\right)+\varepsilon_t}$$



If we assume that the distribution of ε_t is independent of N_t, model 1 is unrealistic. Data, and some theory, speaks in favour of model 3, with ε_t drawn from a normal distribution.

Model fitting

- Guess an appropriate model
  - Based on biology
  - Based on experience
  - Based on the quality and quantity of your data. More and better data allows for a more complex model.
- Adjust parameters to fit data
  - There are a few methods
- Evaluate!
  - Is it realistic?
  - Is it a good fit?
  - Are there alternative models?

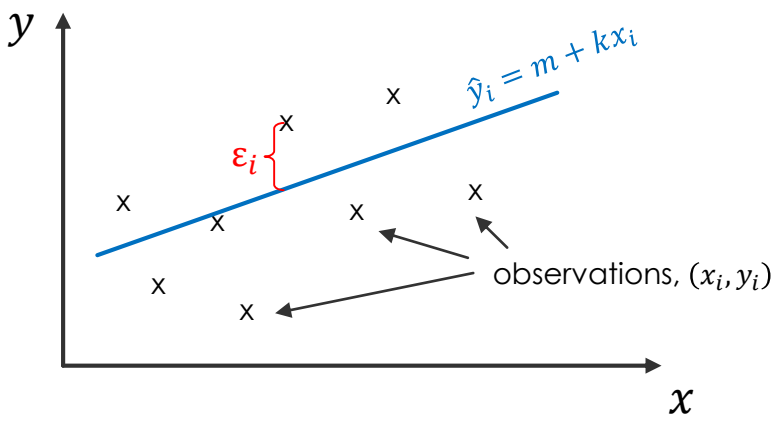
The Least Squares method
Example: Linear Regression

- A common approach to model fitting is to compare model *predictions* to *observations* and make the difference as small as possible.
- Linear regression predicts a variable *y* based on another variable x. The function that relates *y* to x is a straight line, with an intercept and a slope.

Prediction: $\hat{y}_i = m + kx_i$

Observations: (x_i, y_i)

Residuals: *observed – predicted* =
$= \varepsilon_i = y_i - \hat{y}_i$



The Least Squares method
Example: Linear Regression

- The Least Squares method minimizes the *Residual Sum of Squares*, the *RSS*, to find the best values of the model parameters.

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2$$

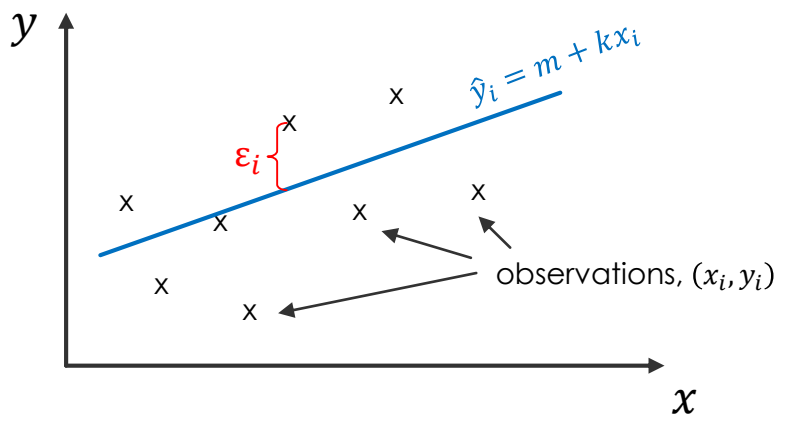
n = number of residuals

- In linear regression the parameters *m* and *k* are adjusted to minimize *RSS*.
- The Least Squares method works best if the residuals follow a normal distribution and are independent of the predictor variable(s).

Prediction: $\hat{y}_i = m + kx_i$

Observations: (x_i, y_i)

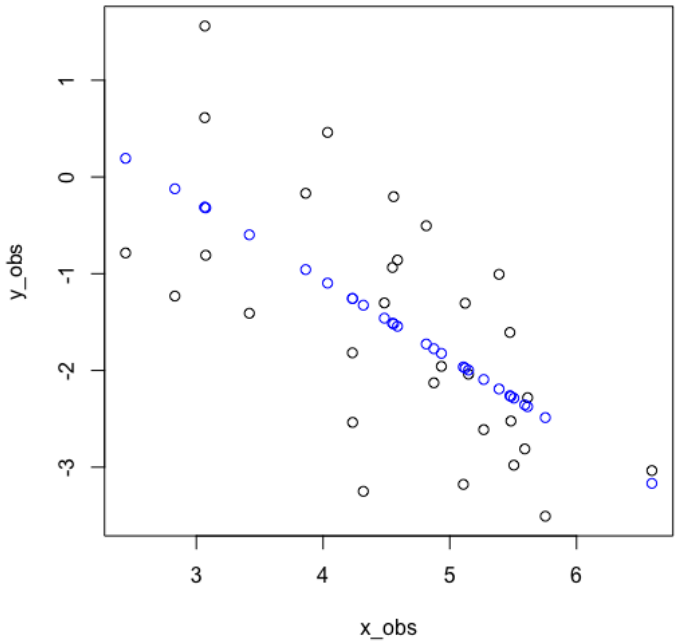
Residuals: *observed – predicted* =
$= \varepsilon_i = y_i - \hat{y}_i$



Using `nls` in R (non-linear least squares)

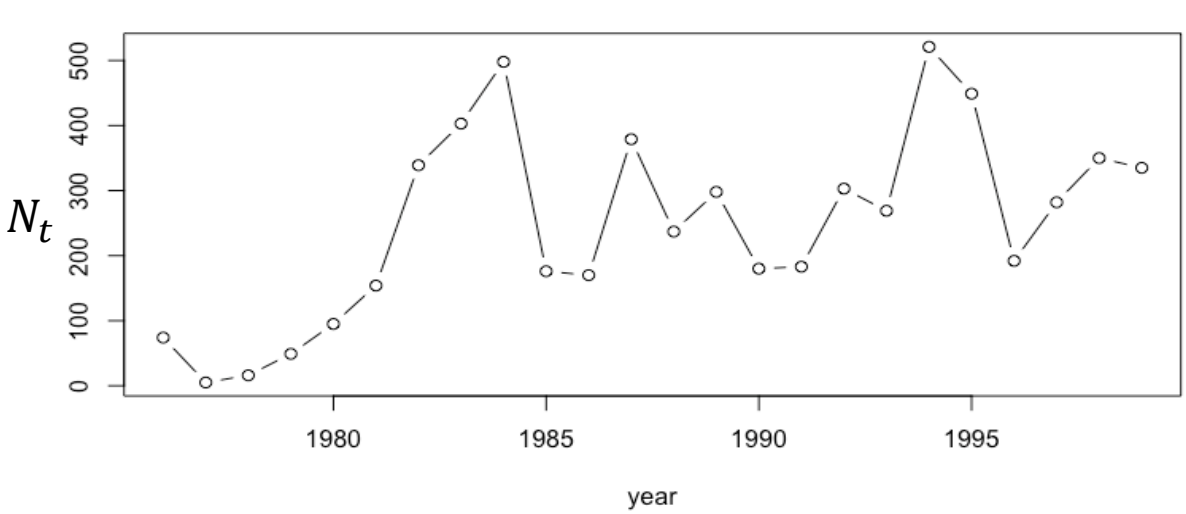
The `nls` function can be used to fit any type of model, not just linear ones, using the least squares method.

```
> x_obs <- rnorm(30) + 5
> y_obs <- 3 - x_obs + rnorm(30)
> plot(x_obs, y_obs)
> fit <- nls( y_obs ~ m + k*x, data=list(x=x_obs), start = list(m=0,k=0))
> summary(fit)

Formula: y_obs ~ m + k * x

Parameters:
  Estimate Std. Error t value Pr(>|t|)
m   2.1679     0.8381   2.587 0.015181 *
k  -0.8090     0.1788  -4.524 0.000102 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9672 on 28 degrees of freedom

Number of iterations to convergence: 1
Achieved convergence tolerance: 2.534e-09

> points(x_obs, fitted(fit), col='blue')
> sum(residuals(fit)^2)
[1] 26.19558
```



Fitting a population dynamic model

- Observations of dynamic systems generate *time series* of the state variable(s).
- A common observation of a population is a time series of its size, density.
- To fit a population model to such data we need to decide on what is *predicted*.



Making predictions

Based on the previous slide, there are at least three ways to make predictions based on the Ricker equation:

1. Predict population size, directly:

$$\widehat{N}_{t+1} = N_t e^{r_0\left(1-\frac{N_t}{K}\right)}$$

observed!

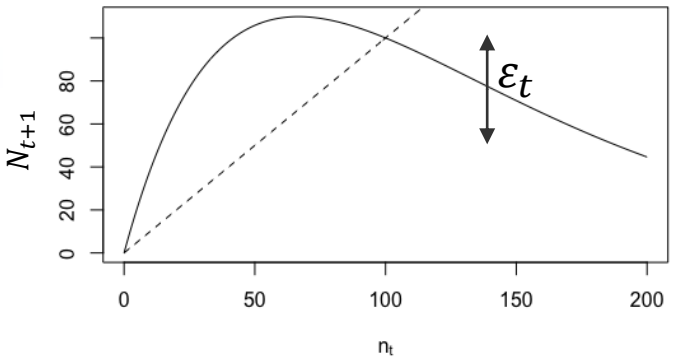
2. Predict the *growth rate*, λ_t:

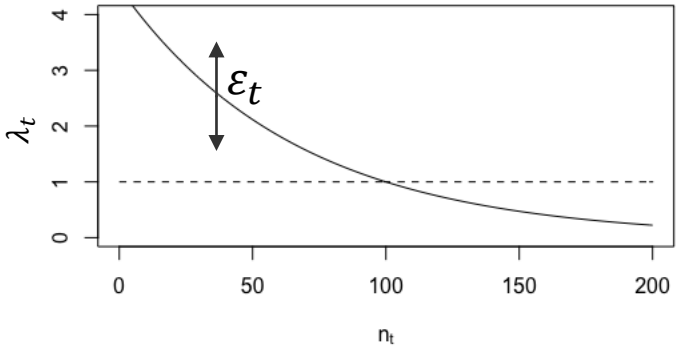
$$\hat{\lambda}_t = \frac{N_{t+1}}{N_t} = e^{r_0\left(1-\frac{N_t}{K}\right)}$$

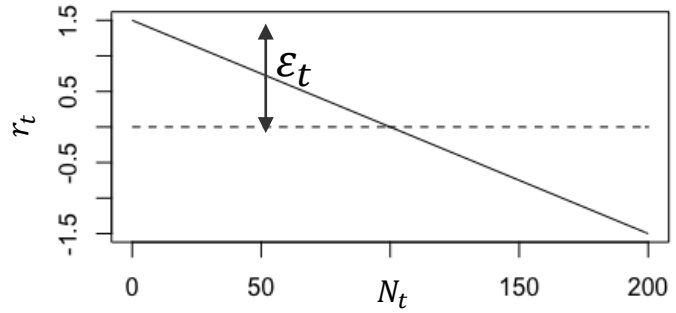
3. Predict the *exponential growth rate*, r_t:

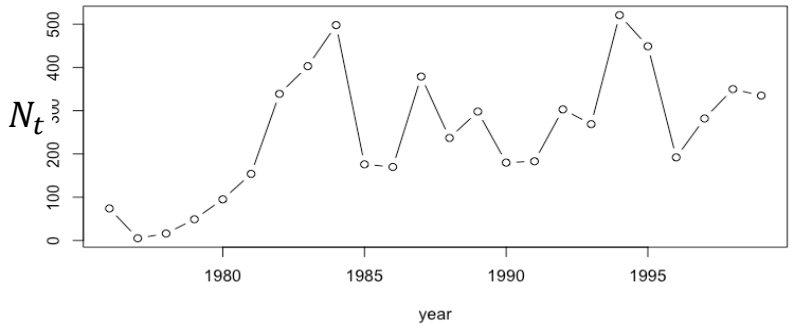
$$\hat{r}_t = \ln(\lambda_t) = r_0\left(1-\frac{N_t}{K}\right)$$

In each case, the predicted value can be compared to the observed values, based on the data.









What is stochastic?

Based on the previous slide, there are at least three ways to add 'stochastic environmental fluctuations' to the Ricker equation:

1. As a random deviation of population size, directly:

$$N_{t+1} = N_t e^{r_0\left(1-\frac{N_t}{K}\right)} + \varepsilon_t$$

2. As a random deviation of the *growth rate*, λ_t:

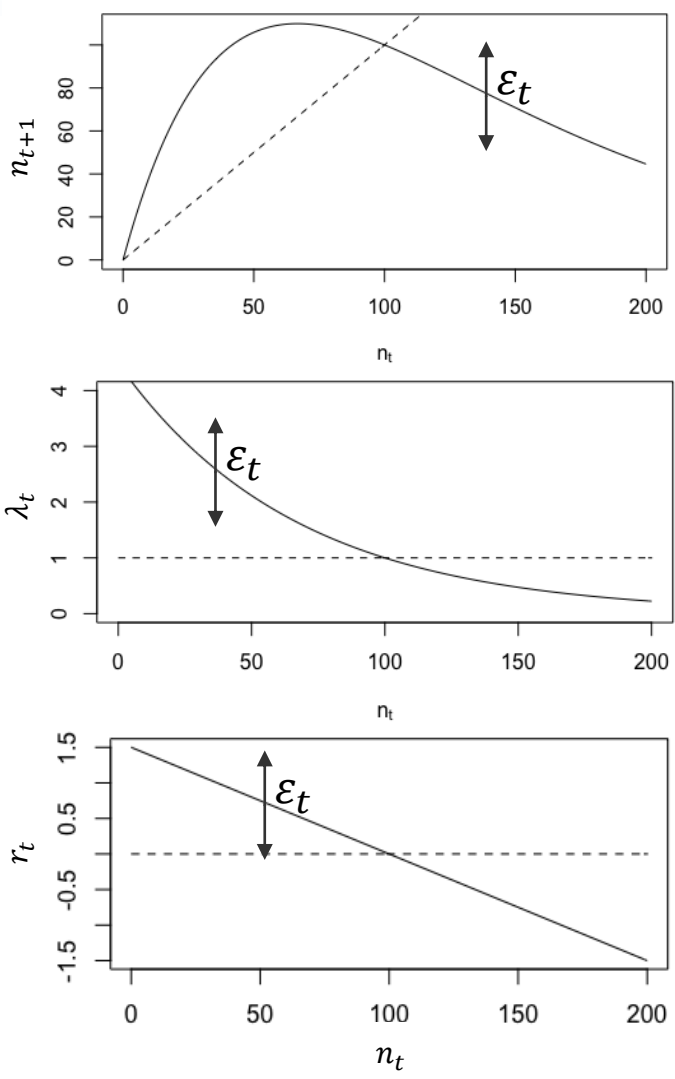
$$\lambda_t = \frac{N_{t+1}}{N_t} = e^{r_0\left(1-\frac{N_t}{K}\right)} + \varepsilon_t$$

$$N_{t+1} = N_t\lambda_t = N_t\left(e^{r_0\left(1-\frac{N_t}{K}\right)} + \varepsilon_t\right)$$

3. As a random deviation of the *exponential growth rate*, r_t:

$$r_t = r_0\left(1-\frac{N_t}{K}\right) + \varepsilon_t$$

$$N_{t+1} = N_t e^{r_t} = N_t e^{r_0\left(1-\frac{N_t}{K}\right)+\varepsilon_t}$$



If we assume that the distribution of ε_t is independent of N_t, model 1 is unrealistic. Data, and some theory, speaks in favour of model 3, with ε_t drawn from a normal distribution.

Model choice

- Often, there are several alternative models that can be used.
- One way to decide which is 'best' is to use the Akaike Information Criterion, the *AIC*

$$AIC = n \ln\left(\frac{RSS}{n}\right) + 2k$$

n = number of residuals

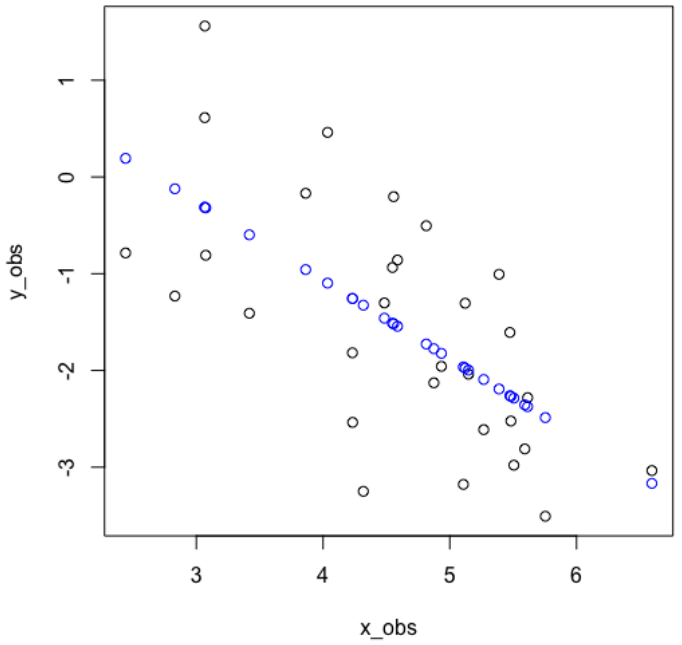
k = number of fitted parameters

- The model with the lowest *AIC* is chosen.
- The *AIC* punishes models with many parameters, which prevents *overfitting*.

Using `nls` in R (non-linear least squares)

The `nls` function can be used to fit any type of model, not just linear ones, using the least squares method.

```
> x_obs <- rnorm(30) + 5
> y_obs <- 3 - x_obs + rnorm(30)
> plot(x_obs, y_obs)
> fit <- nls( y_obs ~ m + k*x, data=list(x=x_obs), start = list(m=0,k=0))
> AIC(fit)
[1] 87.06811
```



Monte-Carlo simulations

- Once we have a fitted model it can be *used* and *interpreted* in various ways.
- *Monte-Carlo* simulations are computer-generated stochastic simulations of a model, used to calculate statistics, such as average outcomes, distributions, probabilities, etc.
- For example, one can calculate the extinction risk of a population.

Monte-Carlo simulations of a fitted population dynamic model

- The `nls` can provide estimates of the r_0 and K parameters of the Ricker equation, fitted to some time series.
- Additionally, we get information about *the size of the residuals*. The standard deviation of the residuals can be estimated as

$$\hat{\sigma} = \sqrt{\frac{RSS}{n-k}}$$

- Assuming the residuals follow a normal distribution, that can be used to generate Monte-Carlo simulations of the population:

$$N_{t+1} = N_t e^{\hat{r}_0\left(1-\frac{N_t}{\hat{K}}\right)+\varepsilon_t}, \qquad \varepsilon_t \in N(0,\hat{\sigma})$$

- In this way, the Monte-Carlo simulations follow the fitted model as close as possible.