Practical 1 - Linear Regression Models

Julien Chiquet & Guillem Rigaill & Anastase Alexandre Charantonis

September 22th, 2016

Session's objectives

- Master Simple Linear Models
- Interpret the `R` outputs of such a model

Remarks

- The lab exercises must be perfomed in groups of two or three students.
- You *must* use `R studio`.
- Using `R studio`, create a `R markdown` file that will detail your work during the lab exercises. That file must be sent by mail to your respective lab monitors, and should include the code and the commentaries on its assosciated outputs. A cheat sheet of the markdown syntax can be found here: https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet
- Whenever possible, prefere the `ggplot2` library to the base plotting functions included in `R`.

For those new to `R`

Go to the page: http://tryr.codeschool.com/. Complete the tutorial.

Quick refresh of `R` basics

1. **Vector Manipulation**. We remind you that

$$e^x = \sum_{k \geq 0} \frac{x^k}{k!}.$$

   Insert in a vector named `exp2` the 20 first elements of this sequence. Remove any values smaller than 10^{-8}. Use the remaining value to give an estimation of e^2 and compare that value to `exp(2)`.

2. **Data Simulation**. Using the `rnorm` function, (type `?rnorm` to get help on this function) generate a vector X containing 100 samples of a normal distribution with a mean value of 2 and a variance of 1. Generate another vector, named Y of the same size obtained by multiplying X by 9.8 and adding a Gausssian noise with a standard deviation of 10.

3. **Read and write a data file**. Put the vectors X and Y in a `data.frame`. Save that `data.frame` with the `write.table` command. Read the obtained table with the `read.table` command. Compare the matrix you obtained to the initial matrix.

4. **Read and write a table in the RData format**. Put the vectors X et Y in a `data.frame`. Save the `data.frame` using the `save` command. Read it with the `load` command. Compare the matrix obtained to the initial one.

5. **Scatter plot**. Trace the scatter plot of Y versus X, first using the `plot` command then using the help of the `ggplot2` library.

6. **Histogram**. Trace the histogram of X.

7. **"For"" loop**. A random variable follows a χ^2 distribution but we do not know its degree of freedom. We want to estimate this degree of freedom. Using the empirical mean of a sample of size n estimate the degree of freedom. Use `R` to evaluate the quality of this estimator for $n = 3$ and $n = 100$. Use a `for` loop.

8. **Alternative "For" loop** (faster, more elegant, more adapted to `R`). Do the same thing with the `sapply` function.

Simple linear regression: Northern pike and DDT

DDT (dichlorodiphenyltrichloroethane) is a relativelly potent pesticide. It of a toxic and non-biodegradable nature and can accumulate in the liver of some adipose tissue. We study here the effect of its accumulation on the Northern pike (*Brochet* in french).

1. Preliminaries

a) Import the data from 'Brochet.txt'.

b) Calculate the mean, median and variance of the age of the Northern pike and of their accumulation of DDT.

c) Plot the histograms of the age of the Northern pike and of their accumulation of DDT.

d) Draw the scatter plot of the accumulation of DDT in the Northern pike versus their age.

e) Make a bloxplot of the accumulation of DDT in the Northern pike versus their age. What do you observe?

f) Calulate the variance of the concentration of DDT in each age group.

2. A first model

a) Create a linear regression model to explain the concentration of DDT in a Northern pike based on its age.

b) Use `R` to estimate the parameters of this model. First apply the formulation given in the lecture, then use the `lm` function. Calculate the slope, bias and residual variance.

c) Test the parameters of the model. Make a variance analysis. Calculate `manually` the Fisher score and the r^2 coefficient.

d) Plot the linear regression. Add confidence intervals. Calculate those first by applying the formulas from in the lecture and then by using the `predict`command.

e) Plot a graph of the residues to evaluate the pertience of your model and perform a diagnostics. You can also use the `plot` function of `R` applied to the outputs of the `lm` function.

3. Square function Modeling

a) Create a new linear regression model to explain the concentration of DDT in a Northern pike based on the square of its age.

b) Use `R` to estimate the parameters of this model.

c) Test the model's parameters and perform a analysis of variance.

d) Plot the regression line. You can use the `geom_smooth` function of `ggplot`.

e) Perform a diagnostic of the model.

4. Logarithmic transformation model

a) Create a new linear regression model to explain the log of the concentration of DDT in a Northern pike based on its age.

b) Trace the box-plot of the log of the DTT concentration in Northern pike as a function of its age. What do you notice?

c) Calculate the variance of the **log** of the DDT concentration in each age group.

d) Use `R` to estimate the parameters of this model.

e) Test the model's parameters and perform an analysis of variance.

f) Plot the regression line. You can use the `geom_smooth` function of `ggplot`.

g) Perform a model diagnostic.

5. Towards a multi-linear regression

a) Create a new linear regression model to explain the log of the concentration of DDT in a Northern pike based on its age and the square of its age.

b) Use `R` to estimate the parameters of this model.

c) Test the parameters of the model. Perform an analysis of variance to compare the 3 models M2, M3, M4 (log, + Age, + square of the Age).

d) Plot the regression curve.

e) Validate your hypotheses. Use `R` to evaluate the relevance of the model. What are your thoughts?

For fast students : numerical maximization of likelihood

Let us consider a phenomenon that can be modeled by a normal distribution: $\mathcal{N}(\mu, \sigma^2)$.

1. Analytically calculate $\log L(x_1, \ldots, x_n; \mu, \sigma^2)$.
2. Determine the maximum likelihood estimators by successivelly deriving $\log L$ with respect to μ and σ^2.
3. Simulate a gaussian sample of size $n = 100$ with a mean of $\mu = \pi/2$ and a standard deviation of $\sigma = \sqrt{2}$. Calculate the values of the maximum likelihood estimators of μ and σ^2 obtained in the previous questions.
4. Create a `loglikelihood` function that takes as inputs `x,mu,sigma` and returns the value of the log-likelihood given $(x_1, \ldots, x_n), \mu$ and σ.
5. Using the `optimize` function, numericaly estimate the values of μ and σ that maximize the `loglikelihood` function.
6. In the same figure, plot the
   - histogram of the data,
   - $\log L$ function for a fixed σ equal to its true value, while homogenously sampling μ in the $[\pi/2 - \varepsilon, \pi/2 + \varepsilon]$ interval; also include the values that were analytically and numerically estimated (use `abline`).
   - $\log L$ function for a fixed μ, while homogenously sampling σ in the $[\sqrt{2} - \varepsilon, \sqrt{2} + \varepsilon]$ interval; also include the values that were analytically and numerically estimated (use `abline`). Add the corrected empirical variance.
7. Create a `logL` matrix, containing the values of the log-likelyhood by simultaneously varying μ and σ. Create a list `data=list(x,y,z)` and use `persp`, `contour`, `image` to visulize the results in 3D and 2D.