DISI – University of Trento

Master in Computer Science AA 2015/2016

Simulation and Performance Evaluation

Assignment 1 (11 points available)

Identification of the revenues vs. downloads characteristics of several web objects

Renato Lo Cigno, Michele Segata, Luca Baldesi

October 29, 2015

A class of web objects (say advertisements linked to browsing generic pages) are associated to two metrics: $x_i(t)$ is the number of downloads of object i at time t, while $y_i(t)$ is the revenue generated (in some way we're not concerned with) by the same object at the same time. Clearly both $x_i(t)$ and $y_i(t)$ are non-decreasing function of t, and indeed, since we assume that the revenue produced is a function of the number of downloads, we can also derive the function $y_i(x_i)$ that represents the revenues generated by object i as a function of the number of its downloads, and we expect also this function to be non-decreasing for each object, though obviously each object can have a different revenue function. As $x_i(t)$ and $y_i(t)$ are sampled at the same time we can consider t a discrete time $t \in \{1, 2, 3, \ldots\}$. Just to fix ideas (but it is completely irrelevant for our problem), suppose that $x_i(t)$ and $y_i(t)$ are generated every hour, so that we have 24 samples per day.

Unfortunately, we do not have access to the accounting unit that records y_i, so that the measures of y_i are affected by noise. However, we know that all the functions $y_i(x_i)$ are expected to show a very simple linear trend

$$y_i(t) = \alpha_i x_i(t) \tag{1}$$

Moreover we know, or better we make the hypothesis, that the increments of $x_i(t)$ are uniformly distributed between 0 (the number of downloads cannot decrease) and a maximum value $D_i^{\max}$, so that

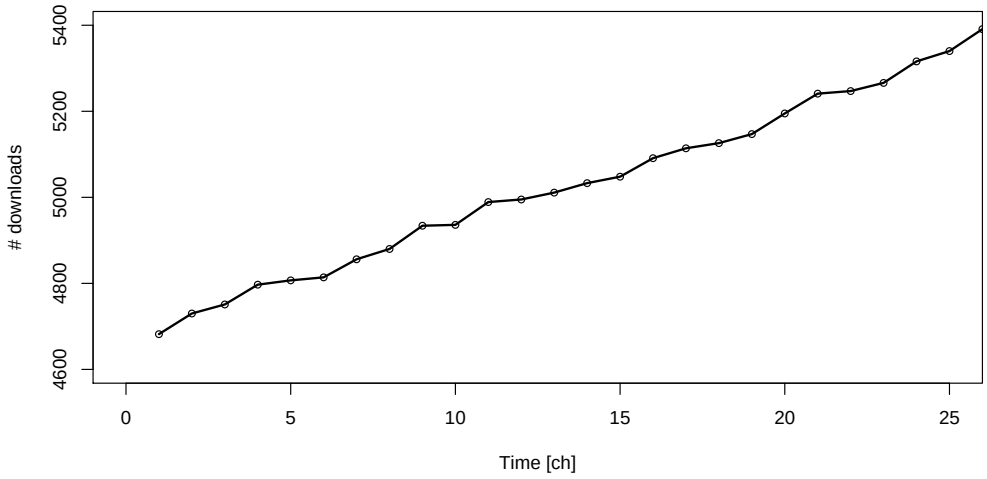
$$\begin{cases} x_i(1) & = D_i \\ x_i(t) & = x_i(t-1) + U(0, D_i^{\max}) \end{cases} \tag{2}$$

where D_i is an initial value known and given for each object and $U(0, D_i^{\max})$ is a **discrete** uniform random variable which generates numbers in the interval $[0, D_i^{\max}]$. We also know that the noise affecting the estimates of the revenues is normally distributed with standard deviation σ and it is independent from the observed object:

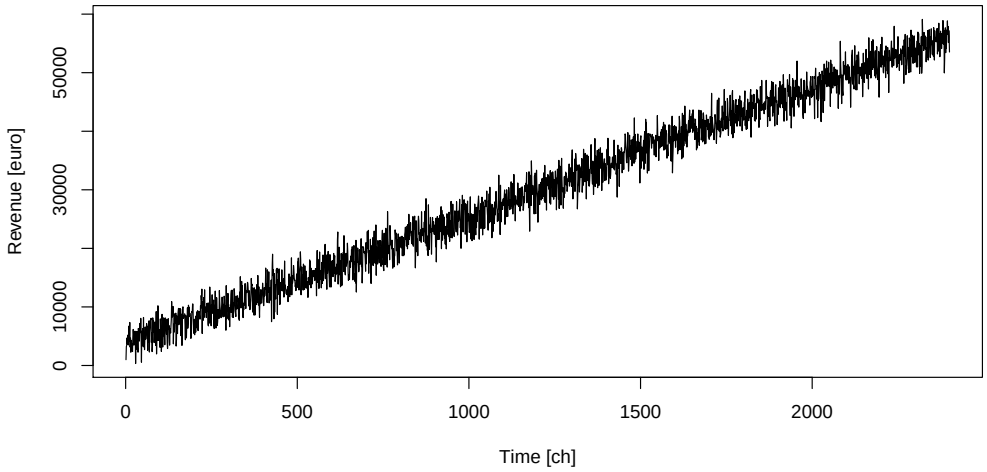
$$y_i(t) = \alpha_i x_i(t) + N(0, \sigma) \tag{3}$$

Fig. 1 plots the unfiltered measures $x(t)$, $y(t)$ and $y(x)$ for a sample object to give a visual representation of the type of data we work with.

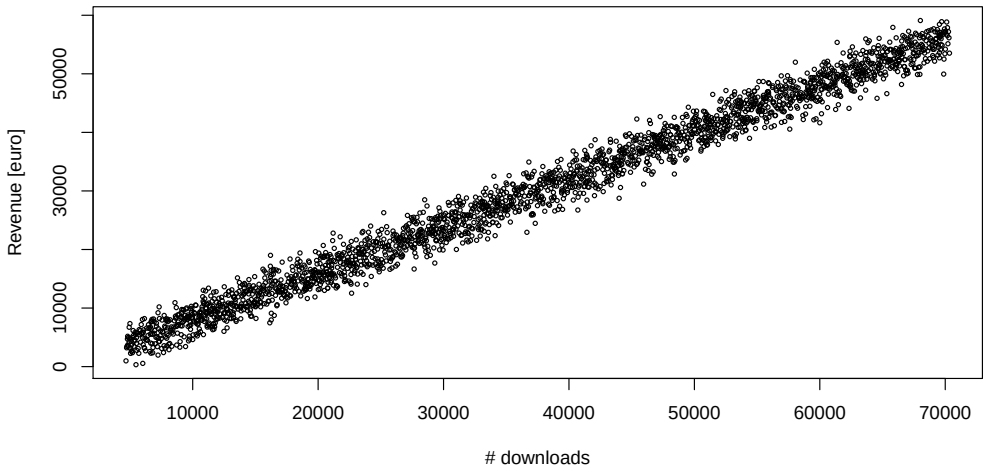
Each object i is "observed" for 100 days taking a measure of $x_i(t)$ and $y_i(t)$ every hour, so that each object is characterized by exactly 2400 measures. Each of you is assigned 10 different object traces in the form of a pair of coordinates (x_i, y_i). Individual assignment files given as csv (comma separated values) are on Classroom. The file has four columns, t, x, y, and i. Each row represents measurement point (x, y) (download, revenues) taken at time t for object i.



(a) downloads vs. time



(b) revenue vs. time



(c) revenue vs. downloads

Figure 1: Plots of the samples x and y vs. t and of the estimation of revenues (y) versus the number of downloads (x) for a generic web object of the class we consider in this assignment

The goal is estimating the 10 sets of parameters $(\alpha_i, D_i^{\max})$, $i = 1, 2, \ldots, 10$ and plot the 10 interpolation polynomials on a (single) x, y plot. Moreover (and obviously), we also want to estimate the parameter σ of the noise affecting the measure of the revenues so that, for instance, new objects can be tracked more easily evaluating if their "trajectory" is promising (high revenues) or not. Moreover you are required to compute the confidence intervals of your estimation of σ. Also, analyzing and commenting on the stochastic processes $x_i(t)$ can help understanding the problem at stake and give insight on how complex processes arise in real life also from simple phenomena. It is up to you to decide how to do this and whether you use t-Student distributions for the errors or a Gaussian approximation, but you have to explain what you do and why, though you can use R or any other tool to do it ... so don't complain about lengthy implementations.

You can work together (indeed, we invite you to do so in pairs or small groups so you can discuss and find better solutions), but remember we will ask you questions about the assignment at the exam, so simple copy-paste without understanding won't be enough ... and the parameters to estimate are different for each student. You will need to use a mathematical tool to estimate the parameters from the samples. You can use the tool you prefer, but we think `R` is a good tool for statistical processing.

Some hints on how to solve the exercise:

- There exists no `findRegressionParametersandSigma` function in statistical tools (as far as we know). You will need to write down some formulas and some "reasoning process" to find a way to solve this, and to explain us how you did. If you find a magic function which solves the problem, you are still required to understand it and explain how it works.
- You can use simple least square error techniques for interpolation (subroutines are available to do this in any statistical analysis tool).
- Think about what you can directly estimate using the samples, but then also consider the stochastic part of the given processes once you subtract the trends.
- Use the LaTex template we gave you and do not write more than 2 pages. Deliver the PDF file of the report and the R or Matlab or whatever scripts you used as a single .zip or .tar file through Classroom.

The deadline to have a correct-and-redo chance for this assignment is December 25 ... 2015!!. If you deliver the assignment within this date, we will correct it and give you the chance to refine it before the exam, otherwise we will consider the work "as is" at the first delivery on Classroom before the oral discussion is agreed upon. However, if the quality of the delivery is unacceptable (e.g., no method is described, plots are meaningless and not explained, etc.) we will not correct it. If you have some doubts, just write us an email or ask at lessons.

Good luck!