
Smoothing and regression splines

Write a report that contains the results of the computations that you are asked to carry out below, as well as the explanation of what you are doing. The main text (2 or 3 pages) should include pieces of source code and graphical and numerical output.
Upload your answers in a .pdf document (use LaTeX or R Markdown, for instance) to ATENEA, as well as the source code (*.R or *.Rmd, for instance). Your work must be reproducible.

The file `bikes.Washington.Rdata` contains information on the bike-sharing rental service in Washington D.C., USA, corresponding to years 2011 and 2012. This file contains only one data frame, `bikes`, with 731 rows (one for each day of years 2011 and 2012, that was a leap year) and 9 columns:

`instant`: row index, going from 1 to 731.

`yr`: year (0: 2011, 1:2012).

`dayyr`: day of the year (from 1 to 365 for 2011, and from 1 to 366 for 2012).

`weekday`: day of the week (0 for Sunday, 1 for Monday, ..., 6 for Saturday).

`workingday`: if day is neither weekend nor holiday is 1, otherwise is 0.

`temp`: temperature in Celsius.

`hum`: humidity in %.

`windspeed`: wind speed in miles per hour.

`cnt`: count of total rental bikes. In this exam we consider this variable as continuous.

1. Consider the nonparametric regression of `cnt` as a function of `instant`. Estimate the regression function $m($`instant`$)$ of `cnt` as a function of `instant` using a cubic regression splines estimated with the R function `smooth.splines` and choosing the smoothing parameter by Generalized Cross Validation.

   a) Which is the value of the chosen penalty parameter λ?

   b) Which is the corresponding equivalent number of degrees of freedom `df`?

   c) How many knots have been used?

   d) Give a graphic with the scatter plot and the estimated regression function $\hat{m}($`instant`$)$.

   e) Estimate now $m($`instant`$)$ by unpenalized regression splines combining the R functions `bs` and `lm`, using the knots
   `my.knots <- quantile(instant,((1:n.knots)-.5)/n.knots)`
   where *n.knots* is the previous value of `df` minus 4.

   f) Give a graphic with the scatter plot and the two estimated regression functions.

2. The script `IRWLS_logistic_regression.R` includes the definition of the function
   `logistic.IRWLS.splines` performing nonparametric logistic regression using splines with a IRWLS procedure. The basic syntax is the following:

   ```
   logistic.IRWLS.splines(x=..., y=..., x.new=..., df=..., plts=TRUE)
   ```

   where the arguments arr the explanatory variable `x`, the 0-1 response variable `y`, the vector `x.new` of new values of variable `x` where we want to predict the probability of `y` being 1 given that `x` is equal to `x.new`, the eauivalent number of parameters (or model degrees of freedom) `df`, and the logical `plts` indicating if plots are desired or not.

   Define a new variable `cnt.5000` taking the value 1 for days such that the number of total rental bikes is larger than or equal to 5000, on 0 otherwise.

   a) Use the function `logistic.IRWLS.splines` to fit the non-parametric binary regression `cnt.5000` as a function of the temperature, using `df=6`. In which range of temperatures is $\Pr($`cnt` $>= 5000 |$ `temp`$)$ larger than 0,5?

   b) **(Optional)** Choose the parameter `df` by k-fold cross validation with $k = 5$ and using `df.v = 3:15` as the set of possible values for `df`.