Advanced Statistical Modeling (ASM). Final Exam, January 2017.

Name: .....

PART 2. Nonparametric modeling: Theory

(1 point) Write about linear smoothers in no more than this page.

Advanced Statistical Modeling (ASM)
Final Exam, January 2017.

PART 2. Nonparametric modeling: Practice

- You can use all the material that you consider appropriate.
- Answers must be given in a pdf file, a Word file, a rtf file or similar.
- In that file you must include everything you consider relevant: **explanations, comments, clarifications**, R instructions, graphics, parts of the outputs provided R, etc. In particular you should be aware of the following:
  - Specify the R libraries and functions that you use in your computation.
  - Indicate what parameters you use in your estimates (in particular the value of the smoothing parameters) and say why you chose those particular values.
- You also must include a file with the R instructions you have used. Your work must be reproducible, so do not forget to include any additional function you had written.
- After finishing the exam, upload your file at **Racó**.

The questions marked with **(*)** require a little bit more elaboration than the other ones (those not marked are very similar to parts of the tasks you have done during the course).

The punctuation of all the questions (including theory and practice) add up to 6 points, even if the maximum grade for this second part of the course is 5 points. You can answer all the questions if you want. In this case I'll discard those items (having total nominal punctuation equal to 1 point) where you obtain the lowest grades. It could be the case that the discarded item is the theretical part of the exam.

The file `bikes.ex.2017.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. (1.5 p.) Consider the nonparametric regression of `cnt` as a function of `instant`.

   a) (0.5 p.) Estimate the regression function $m(\texttt{instant})$ of `cnt` as a function of `instant` using a local linear estimator and choosing the bandwidth by plug-in. Which is the value of the chosen bandwidth? Give a graphic with the scatter plot and the estimated regression function $\hat{m}(\texttt{instant})$.

   b) (0.5 p.) Estimate the conditional variance of `cnt` as a function of `instant`. Draw a graphic with the estimated function $\hat{\sigma}^2(\texttt{instant})$. In a different graphic draw the function $\hat{m}(\texttt{instant})$ and superimpose the bands $\hat{m}(\texttt{instant}) \pm 1{,}96\hat{\sigma}(\texttt{instant})$.

   c) **(*)** (0.5 p.) Test the null hypothesis of constant conditional variance.

2. (1 p.) Now we want to compare the regression function $m(\texttt{dayyr})$ of `cnt` as a function of `dayyr` between years 2011 and 2012.

   a) (0.25 p.) Use library `sm` to test that the regression functions $m(\texttt{dayyr})$ for years 2011 and 2012 are equal.

   b) (0.25 p.) Use library `sm` to test that the regression functions $m(\texttt{dayyr})$ for years 2011 and 2012 are parallel.

   c) **(*)** (0.5 p.) It is possible to test these two hypothesis using `gam` and `anova` from library `mgcv`.

   ```
   library(mgcv)

   gam.2.1 <- gam(cnt ~ s(dayyr))
   summary(gam.2.1)
   plot(gam.2.1,residuals = TRUE, pages=1)

   gam.2.2 <- gam(cnt ~ s(dayyr) + as.factor(yr))
   summary(gam.2.2)
   plot(gam.2.2,residuals = TRUE, pages=1)

   gam.2.3 <- gam(cnt ~ s(dayyr,by=as.factor(yr)) + as.factor(yr))
   summary(gam.2.3)
   plot(gam.2.3,residuals = TRUE, pages=1)
   ```

   Write two calls to function `anova`, one of them to test equality between years and the other one to test parallelism. Justify your answer.

3. (1.5 p.) Use `gam` and `anova` from library `mgcv` to fit several semiparametric models explaining `cnt` as a function of some (or all) of the variables `yr`, `weekday`, `temp`, `hum`, `windspeed` (the first two as factors). Then select the model that you think is the most appropriate. Try to give an interpretation to the selected model in the context of the data set.

4. (1 p.) 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) **(*)** (0.5 p.) 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 $\Pr(\texttt{cnt} >= 5000|\texttt{temp})$ larger than 0,5?

   b) **(*)** (0.5 p.) 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`.