

#### Homework 4. Stat 202a. Due Wed, Nov26, 1159pm.

You must work on the homework INDEPENDENTLY! Collaborating on this homework will be considered cheating. **Late homeworks will not be accepted!** Your homework solution should be a single PDF document. The first pages should be your *output* from the problems. After that, on subsequent pages, include all your *code* for these problems. Submit parts 1 and 2 via the Google Form given to you by your TA. If the Google Form does not work, then email your homework to [statgrader@stat.UCLA.edu](mailto:statgrader@stat.UCLA.edu) . For part 3, email it to [frederic@stat.UCLA.edu](mailto:frederic@stat.UCLA.edu) .

#### 1. Kernel regression and bootstrap standard errors using C.

a) Write a C function to compute the Nadaraya-Watson kernel regression estimate of the relationship between two vectors, X and Y. The inputs to the function should be an integer n, two vectors X and Y each of length n consisting of double-precision numbers, an integer m, a vector g2 of m double-precision gridpoints at which to calculate the kernel regression estimates, a bandwidth b, and a vector res2 of length m which will contain the resulting kernel regression estimates.

b) Gather data on petroleum taxes (X) and consumption (Y) from <https://people.sc.fsu.edu/~jburkardt/datasets/regression/x15.txt>. These are columns 2 and 6 in the dataset.

c) Use your C function to make a kernel regression estimate of the relationship between X and Y, using a Gaussian kernel with bandwidth selected using the rule of thumb suggested by Scott (1992). You may calculate this bandwidth in R. Let  $\{m_1, m_2, \dots, m_{100}\}$  = a vector of 100 equally spaced values spanning the observed range of X in your dataset, compute your kernel regression estimates on this grid using the C function, and plot your kernel regression estimates  $\hat{f}(m_1), \hat{f}(m_2), \dots, \hat{f}(m_{100})$ .

d) In R, sample 100 pairs of observations  $(X_i, Y_i)$  with replacement from your dataset, and for each such sampling, compute another kernel regression estimate,  $\hat{f}(m_1), \hat{f}(m_2), \dots, \hat{f}(m_{100})$ , using the same Gaussian kernel and same bandwidth used in part c).

e) Repeat step d) 200 times, store the results, and for each value of j, find the 2.5th and 97.5th percentile of  $\hat{f}(m_j)$ .

f) Plot your kernel regression estimate from part c) using a solid line, along with the 95% confidence bounds from part e), plotted using dotted lines, on the same plot.

#### 2. Web scraping.

Using either Python or R, scrape data from <https://www.usinflationcalculator.com/inflation/current-inflation-rates/>. Read it into R, and in R, make a plot with months since Jan 2000 on the x-axis, and inflation rate on the y-axis. For example, for Jan 2000, the number of months would be 0, for Feb 2000 it

would be 1, etc. Use `lines()` to connect the dots, and overlay a kernel regression curve as well. You may compute the kernel regression curve simply by using `ksmooth()`, or by using your C function from part 1, whichever you prefer.

### 3. Functions and text manipulation in Python.

Write a function in Python that takes no arguments, and when executed, allows the user to play a game. Your function should start by printing instructions so the user knows what the rules of your game are and how to play. Your game code should consist of a function that, when called, runs the game. If you have other helper functions too, put them lower down, so that the function to play your game is at the very top of your code. Your game should somehow involve scraping some information (such as a random word or random number) from the web. Email your game separately to me at [frederic@stat.ucla.edu](mailto:frederic@stat.ucla.edu). Do not include this as part of your submission on the Google form.

**Output:** Your output for parts 1 and 2 for this assignment should be a pdf document containing the following, in this order.

Figure 1. A plot of your kernel regression estimates  $\hat{f}(m_1)$ ,  $\hat{f}(m_2)$ , ...,  $\hat{f}(m_{100})$  versus  $m$ , along with 95% bootstrap confidence intervals, for the petroleum tax and consumption data.

Figure 2. A scatterplot with months since Jan 2000 on the x-axis, and inflation rate on the y-axis, with the dots connected by line segments and a kernel regression curve overlaid.

After this figure, include all of your C code, R code, and Python code for parts 1 and 2.

Separately, send your python code for part 3 to [frederic@stat.ucla.edu](mailto:frederic@stat.ucla.edu).