Homework 4. Stat 202a. Due Tue, Nov 22, 11pm.

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. Email your homework to statgrader@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 g2of m double-precision gridpoints at which to calculate the kernel regression estimates, 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 $\{m_1, m_2, ..., m_{100}\}$ = 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 $\hat{f}(m_1)$, $\hat{f}(m_2)$, ..., $\hat{f}(m_{100})$.

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, $\hat{f}(m_1)$, $\hat{f}(m_2)$, ..., $\hat{f}(m_{100})$, 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. Python scraping.

Using Python, scrape data from https://en.wikipedia.org/wiki/List_of_most_watched_television_broadcasts_in_the_United_States .
Save the data into a file on your computer, read it into *R*, and in *R*, make a plot with Date on the x-axis, number of viewers on the y-axis, and use different colors to represent the

different networks, for the top 31 tv broadcasts of all time. Ignore #32 which was not associated with a particular network.

Output: Your output 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 of number of viewers (Y) versus date (X), with different colors representing different networks, for the top 31 tv broadcasts of all time.

After this figure, include all of your C code, *R* code, and Python code.