

#### Homework 4. Stat 202a. Due Tue, Nov 27, 10:30am.

You must work on the homework INDEPENDENTLY! Collaborating on this homework will be considered cheating. Submit your homework by email to [stat202a@stat.ucla.edu](mailto:stat202a@stat.ucla.edu). **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 above. After that, on subsequent pages, include all your *code* for these problems.

##### 1. Approximation of an integral in C.

Consider the integral from 0 to  $x_{\max}$  of the shifted Pareto density,  $f(x) = (p-1) c^{p-1} (x+c)^{-p}$ , for  $x \geq 0$ , and  $f(x) = 0$  otherwise, where  $c > 0$  and  $p > 1$  are parameters.

Let  $c = 3$  and  $p = 2$ . Write a C function called *paretoint(xmax,c,p)* that approximates this integral over a grid of 1 million values ranging from  $x = 0$  to  $x_{\max}$ . Call your C function from R to evaluate *paretoint(xmax,c,p)* for various choices of  $x_{\max}$  between 1 and 1000 (you do not need to calculate *paretoint* for every integer between 1 and 1000, but choose around 10-15 numbers between 1 and 1000), and for  $c = 3$  and  $p = 2$  each time. Using R, plot *paretoint(xmax,3,2)* vs.  $x_{\max}$ , for  $x_{\max}$  ranging from 1 up to 1000. You may set up your range of the y-axis in a way that you feel is appropriate.

Repeat the above, but now using  $c = 5$  and  $p = 2.5$ .

##### 2. 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, and a vector res2 of length m which will contain the resulting kernel regression estimates.

b) Gather data on liquor consumption and liver cirrhosis death rates from <http://orion.math.iastate.edu/burkardt/data/regression/x20.txt>. These are the last two columns in the dataset (columns 6 and 7).

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.

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

Figure 1. A plot of  $\text{paretoint}(xmax, 3, 2)$  vs.  $xmax$ , for  $xmax$  ranging from 1 to 1000.

Figure 2. A plot of  $\text{paretoint}(xmax, 5, 2.5)$  vs.  $xmax$ , for  $xmax$  ranging from 1 to 1000.

Figure 3. 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 liquor and cirrhosis death data.

After these 3 figures, include all of your C code, followed by all of your R code.