Stat 624: Final

Complete all problems in a folder in the git repository labeled "Final". Have one R file with all the code for each problem clearly labeled and in order. Name this file MyFinal.R. You will also need an additional file for C or C++ code. Name that file ForFinal.c or ForFinal.cpp. No formal write-up is required.

Problem 1

Suppose you are estimating the parameters in a simple linear regression model where the objective is to make the largest residual as small as possible. Let the model be $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and residuals for a specific model fit are $r_i = y_i - \beta_0 - \beta_1 x_i$. Then if $M(\beta_0, \beta_1) = \max(|r_i|)$, the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimize $M(\beta_0, \beta_1)$.

(a) Write a function that will take data as inputs and return the optimal $\hat{\beta}_0$ and $\hat{\beta}_1$ values found by minimizing $M(\beta_0, \beta_1)$, the maximum residual. You do not need to compute a Hessian and a gradient.

(b) Rewrite the function and expand it to also return a 95% bootstrap confidence interval for β_1.

(c) Perform a simulation study to estimate how often the 95% bootstrap confidence interval contains the true value for β_1. Simulate data using $\beta_0 = -5$, $\beta_1 = 1.8$, and $Var(\varepsilon_i) = 25$. Use an equal number of values for x_i at each of the points 10, 20, 30, and 40, meaning that if there are 20 data points, 5 x's each would equal 10, 20, 30, and 40. Find a Monte Carlo estimate for the coverage of the bootstrap confidence interval for β_1. Repeat this study 3 times using sample sizes of $n = 20$, $n = 100$ and $n = 500$. Report the Monte Carlo estimate along and assess Monte Carlo error of the coverage for each sample size.

(d) Generate an additional data set of size 100 using the parameters in part (c). Plot the data along with two fitted lines: the standard linear regression fitted line and the fitted line found by minimizing $M(\beta_0, \beta_1)$. Make sure these are clearly labeled.

(e) Explore the effect of an outlier on the fitted line. Use the same data set as part (d) but add a value to the generated data set of $(x, y) = (20, 80)$. Fit and plot the two fitted lines. Which of the two lines changes the most?

(f) Determine a clever way to express and present this effect (the effect of an outlier on the fitted line) in a single figure.

Problem 2

Consider the probability density function

$$f(x|\mu, s) = \frac{1}{2s}\left[1 + \cos\left(\frac{x-\mu}{s}\pi\right)\right], \quad \mu - s < x < \mu + s$$

and 0 otherwise.

(a) Write a function in C or C++ that uses a numerical integral to find the k-th moment. From what we talked about in class, using Rcpp is probably the easiest way to do this. State which method of numerical integral you are using (as long as it matches the implementation it can be any we discussed). Be sure to use enough bins to make the numerical method accurate.

(b) Write a function in R that samples from the distribution using Markov chain Monte Carlo methods. Find a proper proposal distribution and tuning parameter to get between 20 and 50% acceptance.

(c) Use the function from part (a) to find the first 4 moments for the distribution using $\mu = 10$ and $s = 5$. Use the function from part (b) to generate samples from the distribution and compare the empirical moments from the samples with the moments calculated using numeric integration. How close are the empirical moments to the ones from the sample?