Homework 2. Stat 202a. Due Tue, Oct 25, 11:59pm. **Late homeworks are not accepted!**

You must work on the homework INDEPENDENTLY! Collaborating with other students on this homework will be considered cheating. 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. Submit your homework by email to statgrader@stat.UCLA.edu .

1. Simulation in R. Kernel density estimates with simulation based 95% confidence bands, applied to earthquake magnitudes and locations.

1a) Gather data on all earthquakes of magnitude at least 4.0 in the longitude range -118.0 to -117.0 and latitude range 34.0 to 35.0, from Jan 1, 1965 to Oct11, 2022, from https://service.scedc.caltech.edu/eq-catalogs/date_mag_loc.php. Input the data into *R*. (Use minimum magnitude = 4.0, maximum magnitude = 9.0, min depth = 0, max depth = 100km, event type = earthquake, geographic type=local.) You will have to input "01" for January and "01" for the day as well. You should have 35 events.

1b) Take the vector of earthquake magnitudes, and use it to make a kernel density estimate of earthquake magnitudes, using a Gaussian kernel with bandwidth selected using the rule of thumb suggested by Scott (1992). Let $\{m_1, m_2, ..., m_{100}\}$ = a vector of 100 equally spaced magnitudes spanning the observed range of magnitudes in your dataset, and plot your kernel density estimates $\hat{f}(m_1)$, $\hat{f}(m_2)$, ..., $\hat{f}(m_{100})$. Use these same values $m_1, m_2, ..., m_{100}$ for parts 1c) and 1d) below.

1c) Simulate 35 earthquake magnitudes drawn independently from your kernel density estimate $\hat{f}$ in part b). Kernel smooth these 35 simulated magnitudes, to produce new kernel density estimates $\hat{f}^*(m_1)$, $\hat{f}^*(m_2)$, ..., $\hat{f}^*(m_{100})$, using the same kernel and bandwidth you used in part 1b).

1d) Repeat step 1c) 200 times. For each value of m_i, find the 2.5th and 97.5th percentiles (i.e., the 5th largest and 195th largest) of your 200 simulated kernel density estimates, $\hat{f}^*(m_i)$.

1e) Extract the longitudes and latitudes of the earthquake origin locations from part 1a), and make a 2-dimensional kernel smoothing of these locations. Overlay the actual locations as points on the plot. Include a legend. Show the kernel-smoothing and locations with a map of (part of) California counties in the background. This might be kind of difficult, especially making the axes of the map agree with the rest of your plot, so do not worry if it does not look perfect.

2. Identification of influential points in multivariate regression.

The purpose of this problem is to investigate how influential each point in a dataset is on a particular regression estimate.

2a) Median sale prices data for Los Angeles County Housing in Aug 2013 from the Los Angeles Times were compiled into the file LAhousingpricesaug2013.txt and placed on the course website. Let Y = sales of single family homes in August, X_1 = median price of a single family residence (SFR) in thousands of dollars, X_2 = median price of a condo in thousands of dollars, and X_3 = median home price per square foot, in dollars. (Note that, in the original LA Times dataset, some of the names of the cities had spaces in them, and some of the numbers have commas in them. These problems have been removed in the LAhousingpricesaug2013.txt file.) Each of these 4 vectors initially has length 269. If any row has an "n/a" in it for any of these 4 variables, then remove this entire row. Now each vector will have length 217.

2b) Perform regression (with intercept) of Y on $X = \{X_1, X_2, X_3\}$ to compute a vector of parameter estimates, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$, where $\hat{\beta}_0$ is the estimated intercept and for i = 1,2,3, $\hat{\beta}_i$ is the slope corresponding to explanatory variable X_i. Record $\hat{\beta}_1$.

2c) Let i = 1. Perform regression with intercept of Y on X with row i removed from the dataset. Let $\hat{\beta}^{(-i)}$ denote your resulting vector of parameter estimates, so that $\hat{\beta}_1^{(-i)}$ is your estimate of the slope corresponding to X_1. Record $\hat{\beta}_1^{(-i)} - \hat{\beta}_1$.

2d) Repeat step 2c) for i = 2, 3, 4, ..., 217.

2e) Plot the influences, $\hat{\beta}_1^{(-i)} - \hat{\beta}_1$, versus i. That is, the x-axis will span from i = 1 to 217, and the y-axis will be $\hat{\beta}_1^{(-i)} - \hat{\beta}_1$, which indicates the influence of observation i on the estimated slope.

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

Figure 1. A plot of your kernel density estimate of the earthquake magnitudes, $\hat{f}$ (m), from part 1b), along with your simulation-based 95% confidence bands from part 1d).

Figure 2. A plot of your 2-dimensional kernel smoothing of the earthquake locations, along with the points themselves and a legend, from part 1e).

Figure 3. A plot of the influences, $\hat{\beta}_1^{(-i)} - \hat{\beta}_1$, versus i, from part 2e).

After these 3 figures, include all of your code.