**Objective**

* Earthquakes are caused by the sudden release of energy
* Initiated at a rupture below the surface
  + initial earthquake= mainshock
  + secondary earthquakes = aftershocks
* We will study a dataset of earthquakes
* We will model the probability of aftershocks
  + based on quantities such as the distance to the mainshock rupture.

**Files Explained**

We have **several tables** with information about earthquakes.

* **all\_events.csv** contains:
  + the date
  + location (latitude and longitude)
  + identifier id
  + intensity mw
  + seismic moment of many earthquakes.
* The tables in the folder aftershocks/ contain:
  + the mechanical stresses s1, ... ,s6 at different locations surrounding a mainshock,
  + a column indicating if an aftershock was identified at that location
    - 0 - aftershock was not recorded
    - 1 - aftershock was recorded
* The table selectedEvents.csv contains:
  + a list of identifiers id
  + a list of the files with the corresponding aftershock tables.

**Questions**:

* Testing:
  + Can we name the file, or should the function be able to accept any file

**Question 1: Data Frames**

1. Question 1a
   1. Create a new **data frame** with four columns:
      1. date = all events
      2. file = selected events
      3. latitude = all events
      4. longitude = all events
      5. mw (intensity) = all events
      6. aftershocks = aftershock file
   2. with a row for each of the selected events containing:
      1. the date (from all\_events.csv),
      2. the file containing the aftershock information (from selectedEvents.csv),
      3. the location of the mainshock,
      4. the intensity
      5. the total number of aftershocks.
   3. Make sure the new data frame is sorted by date,
   4. display the first few rows using head.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Date | File | Latitude | Longitude | Intensity (mw) | Quantity of Aftershocks |
| October 6, 2000 |  |  |  |  |  |
| September 5, 2012 |  |  |  |  |  |
| January 26, 2001 |  |  |  |  |  |
| August 16, 1998 |  |  |  |  |  |

1. Question 1b
   1. Implement a function process\_stress(fi, fu) that receives:
      1. the name of an aftershock file: fi
      2. a function: fu.
         1. receives six arguments (the stress components s1, … ,s6)
         2. returns a single value.
   2. process\_stress returns:
      1. a data frame with columns:
         1. x
         2. y
         3. fu
         4. aftershock
      2. with values from:
         1. the corresponding aftershock file, and
         2. the outputs of the function fu for each row.
   3. Apply it to the event 2001 – BHUJIN 01 YAGI with 𝑓 (𝑠1 , ... , 𝑠6 ) = \_i |s\_i|$, and display the first few rows of the result with head.
2. Question 1c
   1. Create new data frame with four columns:
      1. file (from selectedEvents.csv),
      2. lat,
      3. lon,
      4. moment (from all\_events.csv).
   2. Sort it by the column file and display the first few rows (all of them) with head.

|  |  |  |  |
| --- | --- | --- | --- |
| File | Latitude | Longitude | moment |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Question 2: Mapping**

1. Question 2a
   1. Use geom\_map (Python) and the file worldMap.shp
   2. plot a map of all the events in all\_events.csv
      1. a point for each event.
   3. Note: in Python, read worldMap.shp using geopandas.read\_file.
2. Question 2b
   1. Use geom\_map (Python) with worldMap.shp
   2. plot a map with a point for each event in selectedEvents.csv.
      1. Use colour to represent the intensity,
      2. Use size to represent the number of associated aftershocks.
   3. Note: in Python, read worldMap.shp using geopandas.read\_file.
3. Question 2c
   1. Plot the Euclidean norm of the stresses for 2001 BHUJIN 01 YAGI at the (x,y) coordinates in the corresponding file,
      1. using colour for the value of the norm
      2. include black points at the location of the aftershocks.

**Question 3: Functions**

𝑓(𝛽0, 𝛽1) =

This expression corresponds to the negative log-likelihood of a model. Here 𝑦𝑘 ∈ {0, 1} is the observed outcome (no aftershock or aftershock present), and 𝑥𝑘 is our predictor variable, that we will define based on information about the earthquake.

1. Question 3a
   1. Implement a function fit(X, Y, gamma) that receives:
      1. the vectors with values:
         1. 𝑥𝑘
         2. y𝑘
      2. a step gamma for the gradient descent method,
   2. The function returns:
      1. 𝛽0, 𝛽1
      2. obtained by using the gradient descent method with starting point (0,0).
   3. Test it by computing the values for TODO
      1. Use the Euclidean norm of the stresses as X
      2. Use the value of the column aftershock as Y.
2. Question 3b
   1. Implement a function fit\_file(fi, fu, gamma) that:
      1. finds the optimal values of 𝛽0, 𝛽1
         1. Use gradient descent as before,
      2. Use the data in:
         1. the aftershock file fi,
         2. the function fu on the stresses (defined as in Question 1b).
   2. Test it by computing the values for TODO using the Euclidean norm of the stresses as X and the value of the column aftershock as Y.
3. Question 3c
   1. Implement a function factory fit\_file\_factory(fu, gamma) to:
      1. fix the values of fu and gamma in fit\_file.
   2. Compute the values of 𝛽0, 𝛽1 for all events in 𝑠𝑒𝑙𝑒𝑐𝑡𝑒𝑑𝐸𝑣𝑒𝑛𝑡𝑠, using 𝑓(𝑠1, ... , 𝑠6) = log(∑𝑖 |𝑠𝑖|) and 𝑔𝑎𝑚𝑚𝑎 = 10−3.
   3. Plot the results with 𝛽0 in the 𝑥-axis and 𝛽1 in the 𝑦-axis, one point for each event.

**Question 4: Functions**

The logistic regression model from Question 3 can be extended to more variables, by defining the probability:

1. Question 4a
   1. Write a function moment\_distance(fi) that:
      1. receives the name of an aftershock file, and
      2. returns a data frame with three columns:
         1. the mainshock seismic log-moment (log of moment in all\_events.csv),
         2. the distance between the mainshock
         3. the possible aftershock location computed (assume that the mainshock is at the centre of the grid of points in the aftershock file),
         4. the presence/abscence of an aftershock.
      3. Use the column names moment, distance, aftershock, and
      4. note that the moment is the same for all the rows since we are looking only at one mainshock event.
   2. Display the first few rows of the data frame obtained by applying this function to TODO.
2. Question 4b
   1. Implement a function fit2(X1,X2,Y) that:
      1. minimises the negative log-likelihood function f in Question 3 and returns the values of 𝛽0, 𝛽1 ,𝛽2
   2. Use scipy.optimize.minimize in Python, and do not use the derivative of f.
   3. Obtain the values of 𝛽0, 𝛽1 ,𝛽2 for 2001 BHUJIN 01 YAGI using moment for x1, distance for x2 and aftershock for y.
3. Question 4c
   1. Implement a function fit2\_file(fi) that returns the values of 𝛽0, 𝛽1, 𝛽2 for the aftershock file fi using:
      1. moment for x1,
      2. distance for x2
      3. aftershock for y.
   2. Plot the values of 𝛽1 vs 𝛽0 and 𝛽2 vs 𝛽0 in two separate plots, one point for each event in selectedEvents.csv.