Due: 31st, March 2022, 11am

The worksheet is based on Section 23 in the book R For Data Science. You may refer to the book but write your own `code` and do NOT use `tibble`.

1. Consider the simulated data set `sim1` in `modelr` library.

   (a) Using `ggplot` provide a scatter plot of the `sim$y` versus `sim1$x`.

   (b) Assume that $m \sim \text{Uniform}(-5, 5)$ and $c \sim \text{Uniform}(-20, 40)$. Generate 100 lines with slopes m and intercept c. Plot all the lines layered on top of the scatter plot done above.

   (c) Using the below `function`

   ```
   > RSS = function(a, data) {
   +   d = data$y - (a[1] + data$x * a[2])
   +   sum(d^2)
   + }
   ```

   compute the residual sum of squares for each of the lines.

   (d) Using `ggplot`, the inbuilt function `rank` and `filter` plot the 10 best lines (i.e. 10 lowest RSS) along with the data points. Colour the BRL:=best random line in `viridis plasma red`.

   (e) Understand `optim` function and the command

   ```
   > lsfit=optim(c(0, 0), RSS, data = sim1)
   ```

   Describe the output of the code decide what `lsfit$par` provide and call this BOL:=best optim line.

   (f) Use the inbuilt `lm` function to compute the slope and intercept of least square line and the line LSL:= least square line.

   (g) For LSL, BOL,BRL compute the residuals using the function given below

   ```
   > Residual = function(a, data) {
   +   d = data$y - (a[1] + data$x * a[2])
   +   d
   + }
   ```

   and provide three plots of the same as a histogram and scatter plot.

2. Biologists use a technique called "capture-recapture" to estimate the size of the population of a species that cannot be directly counted. The following exercise illustrates the role a hypergeometric distribution plays in such an estimate.

   (a) Suppose there is a species of unknown population size N. Suppose fifty members of the species are selected and given an identifying mark. Sometime later a sample of size twenty is taken from the population and it is found that four of the twenty were previously marked.[1]

      i. N be the number of population in the wild. Write down the likelihood function for N given the above data.

      ii. Plot the likelihood function for N.

      iii. Use the `optimize` function in `R` to find the maximum likelihood estimate for N.

      iv. Can you compute the M.L.E. for N using calculus ?

[1] The basic idea behind mark-recapture is that since the sample showed $\frac{4}{20} = 20\%$ marked members, that should also be a good estimate for the fraction of marked members of the species as a whole. However, for the whole species that fraction is $\frac{50}{N}$ which provides a population estimate of $N \approx 250$.