**Practical 6**

**Title**

Introduction to linear regression using python and Weka

**Aim**

To learn to perform linear regression using python

**References**

https://stattrek.com/regression/regression-example.aspx

<https://www.geeksforgeeks.org/linear-regression-python-implementation/>

<https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9>

https://www.projectguru.in/publications/procedure-interpretation-linear-regression-analysis-using-stata/

**Perform the following tasks:**

| **Air Velocity (cm/sec)** | 20,60,100,140,180,220,260,300,340,380 |
| --- | --- |
| **Evaporation Coefficient**  **(mm2 /sec)** | 0.18, 0.37, 0.35, 0.78, 0.56, 0.75, 1.18, 1.36, 1.17, 1.65 |

1. For the data in the table given above, compute the estimates for the linear regression coefficient estimates **manually** using the formulas given to you. Calculate R-squared value. Find the value evaporation coefficient for air velocity =240.

The regression equation is a linear equation of the form: ŷ = b0 + b1x . To conduct a regression analysis, we need to solve for b0 and b1. Computations are shown below.

* Find N, sum and mean of each column
* Find (xi - xmean) (yi - ymean), and [ (xi - x)2
* Find b1 = Σ [ (xi - x)(yi - y) ] / Σ [ (xi - x)2]
* Find b0 = y - b1 \* x
* Put the values in formula ŷ = b0 + b1x
* Find σx = sqrt [ Σ ( xi - x )2 / N ], σy = sqrt [ Σ ( yi - y )2 / N ]
* Find R2 = { ( 1 / N ) \* Σ [ (xi - x) \* (yi - y) ] / (σx \* σy ) }2

1. Perform linear regression using Python and Weka
2. . Draw a scatter plot for the data? Does there appear to be a linear relation?
3. . Perform linear regression given the data . Answer the following questions: ( python and weka)
   1. What command did you use to perform the regression? ( python and weka)
   2. What command did you use to view the results of the regression?
   3. Write the regression formula that was obtained.
   4. Is the x-coefficient significant?
   5. Is the constant coefficient significant?
   6. What the residual standard-error value? What is the significance of this value?
   7. What is the R-squared value? What is the significance of this value?
   8. Find the correlation coefficient for this data? Which command did you use? What is the   
      significance of the correlation value?
   9. What is the significance of the F-statistic?
   10. How will you obtain the fitted values for each x-value? Write down the fitted values for each x-value.
   11. How will you obtain the residual values for each x-value? Write down the residual values.
   12. Use a Quantile-Quantile plot to determine if the residuals are normally distributed? Write   
        down your evaluation of the Quantile-Quantile plot.
4. . Perform the following tasks:
   1. Load the 'baseball.arff' file in weka.
   2. Peform linear regression on x:bat\_ave vs y:homeruns and note down the linear regression equation and other relevant values.
   3. Create a Quantile-Quantile plot of the residuals? Are the residuals normally distributed?
   4. Perform a log transformation on the 'homeruns' colum, perform linear regression again,   
       and note down the linear regression equation and all relevant values.
   5. Create a Quantile-Quantile plot of the residuals? Are the residuals normally distributed?