**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 |

* 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
* Perform linear regression using Python and Weka
* **. Draw a scatter plot for the data? Does there appear to be a linear relation?**

There is a linear relation.



* . Perform linear regression given the data . Answer the following questions: ( python and weka)
* **What command did you use to perform the regression? ( python and weka)**

weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

* **What command did you use to view the results of the regression?**
* **Write the regression formula that was obtained.**

y = 14.8571 + 188X

* **Is the x-coefficient significant?**

Yes, the x-coefficient is significant because it is affecting the linear regression decision.

* **Is the constant coefficient significant?**

Yes, the constant coefficient is significant because it is affecting the linear regression decision

* **What the residual standard-error value? What is the significance of this value?**

The residual standard-error value is 5.137400784702911.

The residual standard deviation is a statistical term used to describe the difference in standard deviations of observed values versus predicted values as shown by points in a regression analysis.

* **What is the R-squared value? What is the significance of this value?**

0.97116

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

* **Find the correlation coefficient for this data? Which command did you use? What is the significance of the correlation value?**

corr = (float)(n \* sum\_XY - sum\_X \* sum\_Y)/(float)(math.sqrt((n \* squareSum\_X - sum\_X \* sum\_X)\* (n \* squareSum\_Y - sum\_Y \* sum\_Y)))

The correlation is 0.9952.

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement.

* **What is the significance of the F-statistic?**

The *F* statistic is 51.5105.

The *F* statistic is the ratio of a measure of the variation in the group means to a similar measure of the variation within the groups. If the null hypothesis is correct, then the numerator should be small compared to the denominator. A small *F* statistic will result, and the area under the *F* curve to the right will be large, representing a large *p*-value. When the null hypothesis of equal group means is incorrect, then the numerator should be large compared to the denominator, giving a large *F* statistic and a small area (small *p*-value) to the right of the statistic under the *F* curve.

* **How will you obtain the fitted values for each x-value? Write down the fitted values for each x-value.**

y\_test\_predict = lin\_model.predict(X\_test)

Fitted x-values are 37.38999403, 29.79290611, 25.86755297, 0.31370828, 33.31385559

* How will you obtain the residual values for each x-value? Write down the residual values.

Formula to calculate residuals X\_test - y\_test\_predict

* Use a Quantile-Quantile plot to determine if the residuals are normally distributed? Write down your evaluation of the Quantile-Quantile plot.
* . Perform the following tasks:
* Load the 'baseball.arff' file in weka.
* Peform linear regression on x:bat\_ave vs y:homeruns and note down the linear regression equation and other relevant values.
* Create a Quantile-Quantile plot of the residuals? Are the residuals normally distributed?

Linear Regression Model

1987\_average\_salary =

59193.9498 \* league=N +

90758.9768 \* division=E +

-37248.2691 \* position\_in\_final\_league\_standings\_in\_1986=4,3,2,5,1 +

62364.2195 \* position\_in\_final\_league\_standings\_in\_1986=2,5,1 +

-56539.5067 \* position\_in\_final\_league\_standings\_in\_1986=5,1 +

9139.1519 \* position\_in\_final\_league\_standings\_in\_1986=1 +

-134085.7792 \* team=Pit.,Tex.,Mil.,Mon.,Cin.,S.F.,Cle.,Oak.,S.D.,K.C.,Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

179025.1951 \* team=Tex.,Mil.,Mon.,Cin.,S.F.,Cle.,Oak.,S.D.,K.C.,Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

-60945.3681 \* team=Mil.,Mon.,Cin.,S.F.,Cle.,Oak.,S.D.,K.C.,Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

44632.908 \* team=Cin.,S.F.,Cle.,Oak.,S.D.,K.C.,Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

62141.9414 \* team=S.F.,Cle.,Oak.,S.D.,K.C.,Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

34081.7978 \* team=Cle.,Oak.,S.D.,K.C.,Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

125977.6637 \* team=Oak.,S.D.,K.C.,Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

-97360.817 \* team=S.D.,K.C.,Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

25362.1844 \* team=K.C.,Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

55205.3521 \* team=Cal.,Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

-74965.1224 \* team=Tor.,Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

67231.8516 \* team=Hou.,Min.,St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

-126594.7207 \* team=St.L.,Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

91473.78 \* team=Det.,Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

-19304.0715 \* team=Atl.,Chi.,N.Y.,Phi.,Bal.,Bos.,L.A. +

54256.2368 \* team=Bal.,Bos.,L.A. +

-6394.1236 \* number\_of\_wins\_in\_1986 +

-3422.2589 \* number\_of\_losses\_in\_1986 +

0.0402 \* attendance\_for\_home\_games\_in\_1986 +

0.1506 \* attendance\_for\_away\_games\_in\_1986 +

663519.2691

Time taken to build model: 0.15 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.2296

Mean absolute error 107383.4475

Root mean squared error 125997.6583

Relative absolute error 118.1006 %

Root relative squared error 111.7583 %

Total Number of Instances 26

* Perform a log transformation on the 'homeruns' colum, perform linear regression again,   
   and note down the linear regression equation and all relevant values.
* Create a Quantile-Quantile plot of the residuals? Are the residuals normally distributed?