**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 |

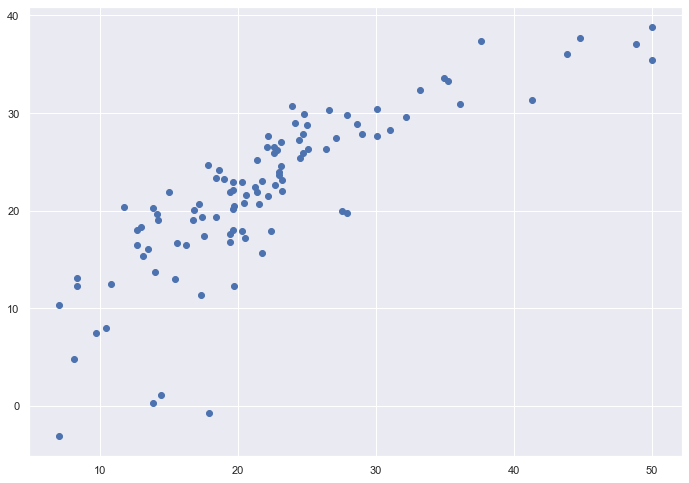
1. For the data in the table given above, compute the estimates for the linear regression coefficient estimates **manually** using the formulas given to you. Calculate R-squared value. Find the value evaporation coefficient for air velocity =240.

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?

There is a linear relation.



1. . 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)

LinearRegression.fit

* 1. What command did you use to view the results of the regression?

**Scikit-learn**

R2\_score

LinearRegression.coef\_ ….{ to see the coefficients}

**Statsmodel**

model.summary

  OLS Regression Results

==============================================================================

Dep. Variable:                  price   R-squared:                       0.994

Model:                            OLS   Adj. R-squared:                  0.965

Method:                 Least Squares   F-statistic:                     33.96

Date:                Wed, 01 Apr 2020   Prob (F-statistic):              0.130

Time:                        16:14:41   Log-Likelihood:                -71.910

No. Observations:                   7   AIC:                             155.8

Df Residuals:                       1   BIC:                             155.5

Df Model:                           5

Covariance Type:            nonrobust

==================================================================================

                     coef    std err          t      P>|t|      [0.025      0.975]

----------------------------------------------------------------------------------

const           -1.74e+04   4.91e+04     -0.354      0.783   -6.41e+05    6.06e+05

size             155.4098     63.389      2.452      0.247    -650.020     960.839

land              36.0683      9.508      3.794      0.164     -84.740     156.876

rooms           1.294e+04   1.61e+04      0.801      0.570   -1.92e+05    2.18e+05

granite         7.954e+04   2.06e+04      3.866      0.161   -1.82e+05    3.41e+05

extra\_bathroom  7.728e+04   1.64e+04      4.718      0.133   -1.31e+05    2.85e+05

==============================================================================

Omnibus:                          nan   Durbin-Watson:                   1.541

Prob(Omnibus):                    nan   Jarque-Bera (JB):                1.062

Skew:                           0.949   Prob(JB):                        0.588

Kurtosis:                       3.200   Cond. No.                     2.77e+04

==============================================================================

* 1. Write the regression formula that was obtained.

price = 195.2035 \* size + 38.9694 \* land + 76218.4642 \* granite + 73947.2118 \* extra\_bathroom + 2681.136

* 1. Is the x-coefficient significant?

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

* 1. Is the constant coefficient significant?

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

* 1. 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.

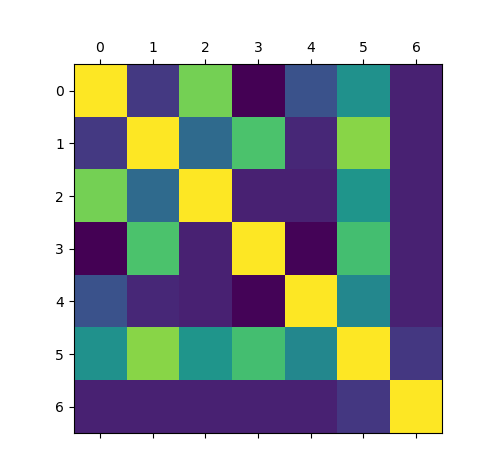
* 1. What is the R-squared value? What is the significance of this value?

0.9952

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.

* 1. 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 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.

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

The F statistic is 33.96.

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.

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

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

Fitted y-values are

[328483.93639079 468595.17201605 427173.47336172 383136.06360921

321456.63715632 604388.76437474 368785.95309117]

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

Formula to calculate residuals y\_test - y\_predict

Residual :

0 -3983.936391

1 -2595.172016

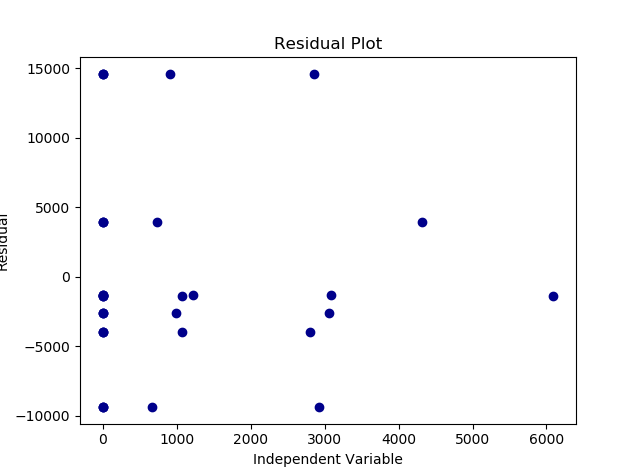
2 -1273.473362

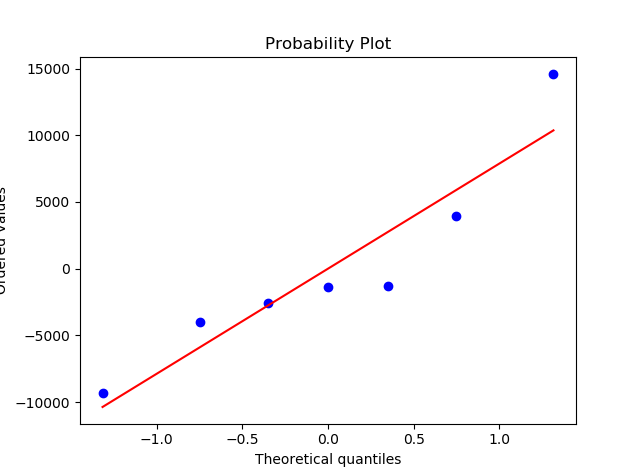
3 3983.936391

4 -9356.637156

5 -1388.764375

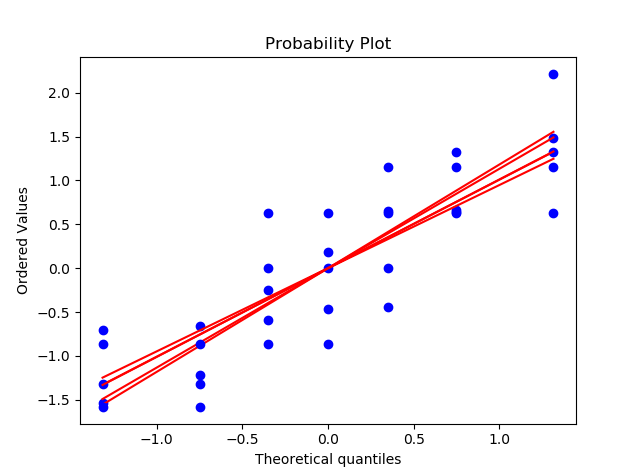
6 14614.046909



* 1. Use a Quantile-Quantile plot to determine if the residuals are normally distributed? Write   
      down your evaluation of the Quantile-Quantile plot. 

Quantile-Quantile of Residual

No the residuals are not normally distributed



Quantile-Quantile of size,land rooms, granite and extra\_bathroom

1. . 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.

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

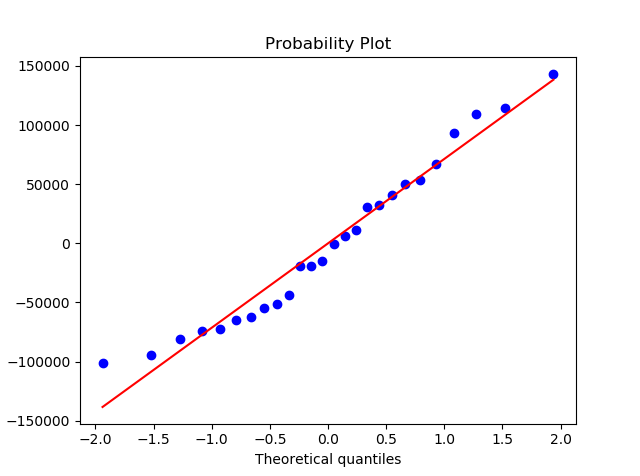
Relative absolute error 118.1006 %

Root relative squared error 111.7583 %

Total Number of Instances 26

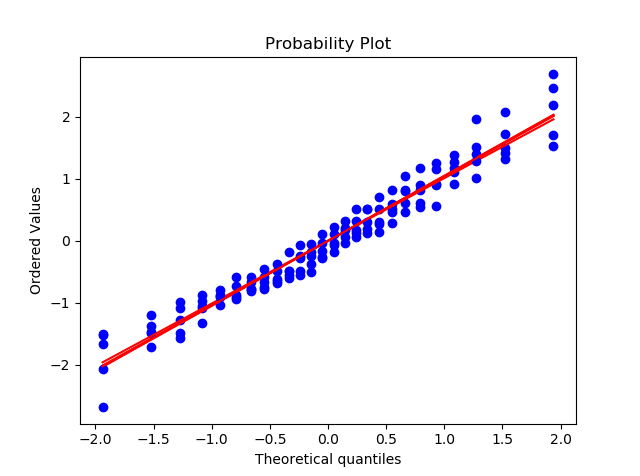
* 1. Create a Quantile-Quantile plot of the residuals? Are the residuals normally distributed?

Linear Regression Model



Quantile quantile for Residuals

No the residuals r not normally distributed



Quantile Quanitle for number\_of\_wins\_1986,numer\_of\_losses\_1986,attendance\_for\_home\_games\_in\_1986, ,attendance\_for\_away\_games\_in\_1986 and 1987\_average\_salary

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