**EAST WEST UNIVERSITY**

**LAB – 6**

**Linear Regression**

**Course Code: ICE470**

**Course Title: Applied Numerical Methods**

**Section – 01**

**Submitted To:**

Muhammad Suhail Najeeb  
Lecturer

Dept. of ECE

East West University

**Submitted By:**

MD Imran Ahmed

ID: 2015-2-50-018

**Date of Submission**: 22-07-2019

**Objective**: Use linear regression to find the best linear fit of the following data

**Ages of developer between 18 to 55**

x = [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55]

**Python Developer Salary by age**

y = [20046, 17100, 20000, 24744, 30500, 37732, 41247, 45372, 48876, 53850, 57287, 63016, 65998, 70003, 70000, 71496, 75370, 83640, 84666, 84392, 78254, 85000, 87038, 91991, 100000, 94796, 97962, 93302, 99240, 102736, 112285, 100771, 104708, 108423, 101407, 112542, 122870, 120000]

**MATLAB Code:**

clc;

clear all;

close all;

%Ages of developer between 18 to 55

x = [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55]

%Python Developer Salary by age

y = [20046, 17100, 20000, 24744, 30500, 37732, 41247, 45372, 48876, 53850, 57287, 63016, 65998, 70003, 70000, 71496, 75370, 83640, 84666, 84392, 78254, 85000, 87038, 91991, 100000, 94796, 97962, 93302, 99240, 102736, 112285, 100771, 104708, 108423, 101407, 112542, 122870, 120000]

n = 38;

scatter(x,y);

grid on;

%plot(x,y)

hold on;

%linear Regression Alogithm

sumx = 0;

sumy = 0;

sumxy = 0;

sumx2 = 0;

st = 0;

sr = 0;

for i = 1:n

sumx = sumx+x(i);

sumy = sumy+y(i);

sumxy = sumxy + x(i) \* y(i);

sumx2 = sumx2 + x(i) \* x(i);

end;

xm = sumx/n

ym = sumy/n;

a1 = (n\*sumxy - sumx\*sumy) / (n\*sumx2 - sumx\*sumx);

a0 = ym - a1\*xm;

fprintf('the equation is: y = %.5f + %.5f x\n',a1,a0);

%Error Analysis of the linear fit:

fprintf('computations for an error analysis of the linear fit: \n');

fprintf('x\t y \t \t \t st \t \t \t sr \n');

for i = 1:n

st0 = (y(i) - ym)^2;

sr0 = (y(i) - a1\*x(i) - a0)^2;

st = st+st0;

sr = sr+sr0;

fprintf('%d \t %.5f \t %.5f \t %.5f \n',x(i),y(i),st0,sr0);

end;

fprintf('--------------------------------------\n');

fprintf('%d \t %.5f \t %.5f \t %.5f \n',sumx,sumy,st,sr);

syx = (sr/(n-2))^(0.5);

r2 = (st - sr)/st;

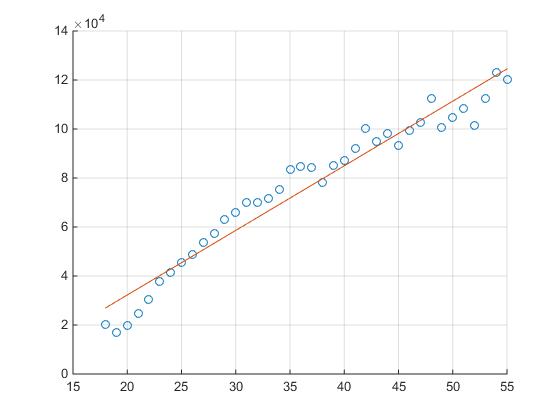
fprintf('sy/x = %.5f & r2 = %.5f',syx,r2);

%plotting the result

yr = a0 +a1.\*x;

plot(x,yr);

**Plot:** Bestfitted line through the data

****

**Output:**

The Linear Equation is: **Y = 2636.89594 + -20492.49130 X**

Computations for an error analysis of the linear fit:

x y st sr

18 20046.00000 3103404720.04432 47964428.84532

19 17100.00000 3440316412.46537 156463361.98778

20 20000.00000 3108531991.41274 149950494.86544

21 24744.00000 2602041577.93906 102785602.35563

22 30500.00000 2047943570.36011 49269440.83695

23 37732.00000 1445688493.30748 5876335.13268

24 41247.00000 1190747578.30748 2390150.84814

25 45372.00000 923078716.46537 3353.24507

26 48876.00000 722438201.09695 654799.53990

27 53850.00000 479794438.78116 9899209.40491

28 57287.00000 341037864.62327 15574112.16925

29 63016.00000 162262007.41274 49540609.32662

30 65998.00000 95183643.83379 54517742.21514

31 70003.00000 33076422.51801 76592553.00877

32 70000.00000 33110938.78116 37354358.45193

33 71496.00000 18132356.88643 24710097.99366

34 75370.00000 147617.72853 38539628.09846

35 83640.00000 62185675.62327 140212439.82707

36 84666.00000 79419991.62327 104657758.15902

37 84392.00000 74611406.99169 53572760.04408

38 78254.00000 6248947.41274 2118638.71915

39 85000.00000 85484622.99169 7041325.59905

40 87038.00000 127323904.88643 4221601.75305

41 91991.00000 263633332.41274 19103523.23529

42 100000.00000 587858307.20222 94923356.07479

43 94796.00000 362589746.36011 3617474.13488

44 97962.00000 493185913.30748 5910100.95652

45 93302.00000 307924915.41274 23676262.86034

46 99240.00000 551582307.20222 2448354.81470

47 102736.00000 728016963.20222 497896.62264

48 112285.00000 1334498579.57064 38520470.43221

49 100771.00000 625839755.57064 63113646.78834

50 104708.00000 838321924.88643 44146798.53706

51 108423.00000 1067249805.67590 30982600.95543

52 101407.00000 658065607.78116 231620931.87023

53 112542.00000 1353341454.36011 45171754.22053

54 122870.00000 2219897617.72853 941114.41204

55 120000.00000 1957689886.14958 20582421.99341

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

1387 2878660.00000 33531907218.31579 1759167510.33549

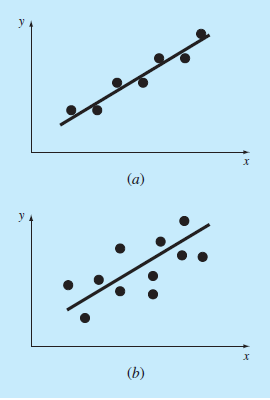
Sy/x = 6990.40515 & r2 = 0.94754>>

**Discussion:**

**Residual Error:**

Sr= =

The difference between the observed value (y) and the predicted value () is called the residual (e). When we look at a regression plot, the residual is the distance from the data point to the fitted regression line.



Above figure (a) has less residual error than the figure (b).

**Standard deviation:**

Standard deviation for the regression line

S(y/x) = = 6990.40515

This is called the *standard error of the estimate*. The subscript notation (y/x) designates that the error is for a predicted value of y corresponding to a particular value of x. also we divide by n-2 because two data-derived estimates a0 and a1 were used to compute Sr

**Coefficient of determination:**

r2 = = 0.94754

R squared, also known as the coefficient of determination, is a measure to indicate how close the data is to the fitted regression line. The value of the R-Squared is the percentage of variation of the response variable(y) that is explained by a linear model. R-squared is always between 0 and 100%. 0% indicates that the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean.

We can say that 94.754% of the variation of the salary is explained by this simple linear model