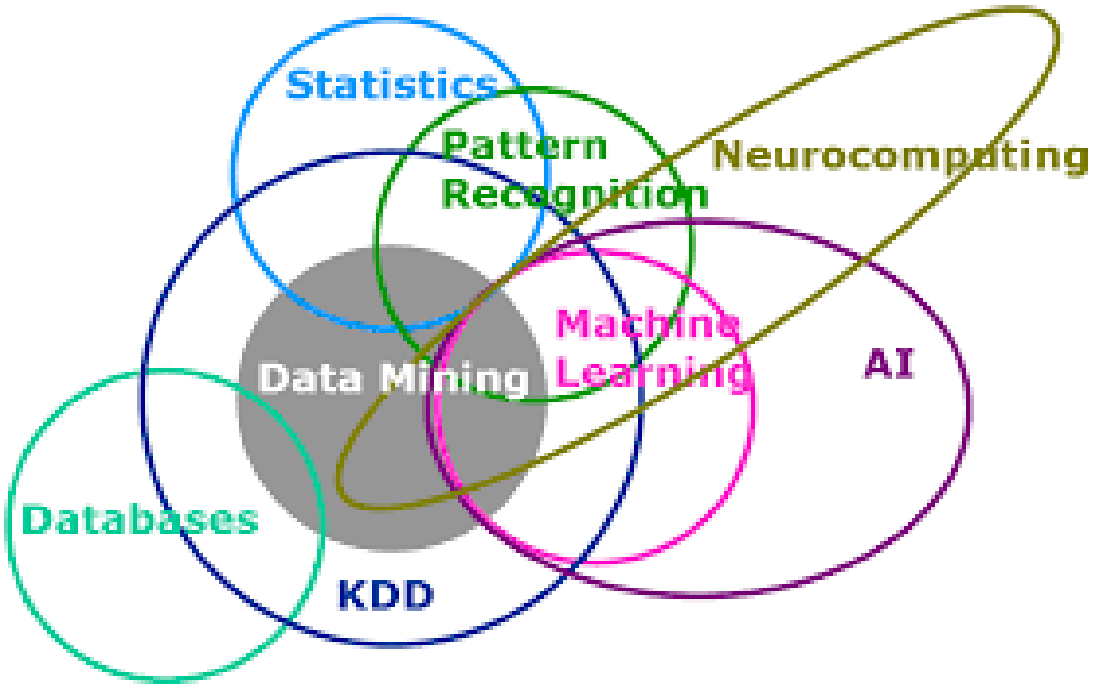
In the name of Allah the most Beneficial ever merciful

بسم الله الرحمن الرحيم



Artificial Intelligence (AI) in Software Engineering

Regression Statistics Table

Department of Computer Science , Univeristy of Karachi (DCS-UBIT)

4th May 2021

Agenda

- ✓ Presentation Group 1
- ✓ Predict the dependent variable ($\hat{Y}$)
- ✓ Estimate the effect of each independent variable (X) on the dependent variable (Y)
- ✓ Calculate the correlation between the dependent variable and the independent variables.
- ✓ Test the linear model significance level.

Groups

GROUP #01			
B18158054	Hasan Haider	Page 2	11th May 2021
B18158014	Zaid khan	Page 3	11th May 2021
B18158067	Zainab	Dataset description, pre-processing , outlier	11th May 2021
B18158056	Shahzain	Feature Extraction	11th May 2021
B18158043	Munazza	Why Model Selection ??	11th May 2021

(1) Ask group 1 to download the JM1 dataset from the given link and prepare a PowerPoint presentation as per the attached plan.
Software Defect Prediction Data Analysis | Kaggle

(2) The Kaggle link can also provide help for the extension and improvement of today's lab data visualization and analysis.
(3) All the students are supposed to try their best to run code from Kaggle as the next week's lab practice.
(4) Your work must show that it belongs to you. Personalization matters in every aspect of life.
(4) Group 1 is supposed to Email their presentation and kaggle lab practice notebook at least before one day of next week's lecture
(5) I will discuss and elaborate on the same topics after students finish with their presentation effort and kaggle lab demonstration.
(6) Group 1 can write their properly formulated queries in case they face difficulty in preparation maximum by Sunday 9th May 202
will be happy if they show courage to handle things in their own way.

UNIVERSITY OF KARACHI

Use case I: Predict the dependent variable ($\hat{Y}$)

Problem Statement

A company sets different LOC rates for a particular project in its eight different modules. The accompanying table shows the numbers of LOC and the corresponding rates.

LOC	420	380	350	400	440	380	450	420
Rates (100USD)	5.5	6.0	6.5	6.0	5.0	6.5	4.5	5.0

What is linear regression?

The linear regression is the linear equation that best fits the points.

There is no one way to choose the best fitting line, the most common one is the ordinary least squares (OLS).

The linear regression describes the relationship between the dependent variable (Y) and the independent variables (X).

The linear regression model calculates the dependent variable (DV) based on the independent variables (IV, predictors).

Quick Linear Regression Calculator (socscistatistics.com)

Linear Regression Calculator

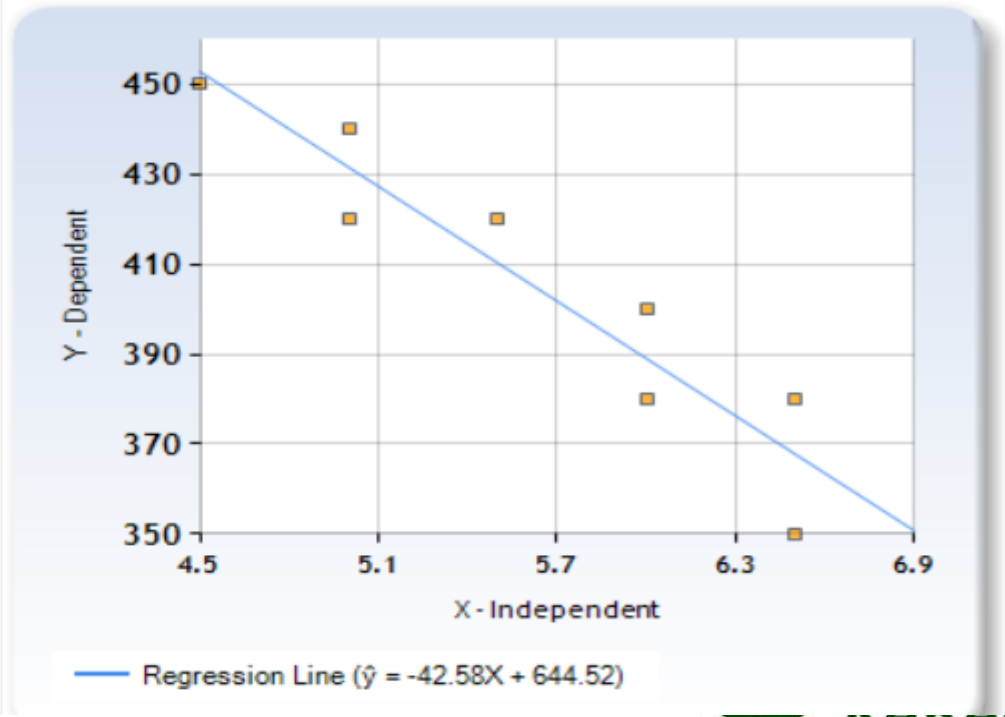
For your data, the regression equation for *Y* is:

$$\hat{y} = -42.58065X + 644.51613$$

As you can see the output from this calculator is fairly verbose. Mostly it should be self-explanatory, but you should note that any apparent discrepancies in calculations are because rounding is used for the purposes of display, but not for the calculations themselves.

If you wish to perform a further calculation, it is necessary to hit the reset button at the bottom of the page.

X Values	*Y* Values
5.5	420
6.0	380
6.5	350
6.0	400
5.0	440
6.5	380
4.5	450
5.0	420
M: 5.625	M: 405



Calculation Summary

Calculation Summary

Sum of $X = 45$
Sum of $Y = 3240$
Mean $X = 5.625$
Mean $Y = 405$
Sum of squares (SS_X) = 3.875
Sum of products (SP) = -165

Regression Equation = $\hat{y} = bX + a$

$b = SP/SS_X = -165/3.88 = -42.58065$

$a = M_Y - bM_X = 405 - (-42.58*5.63) = 644.51613$

$\hat{y} = -42.58065X + 644.51613$

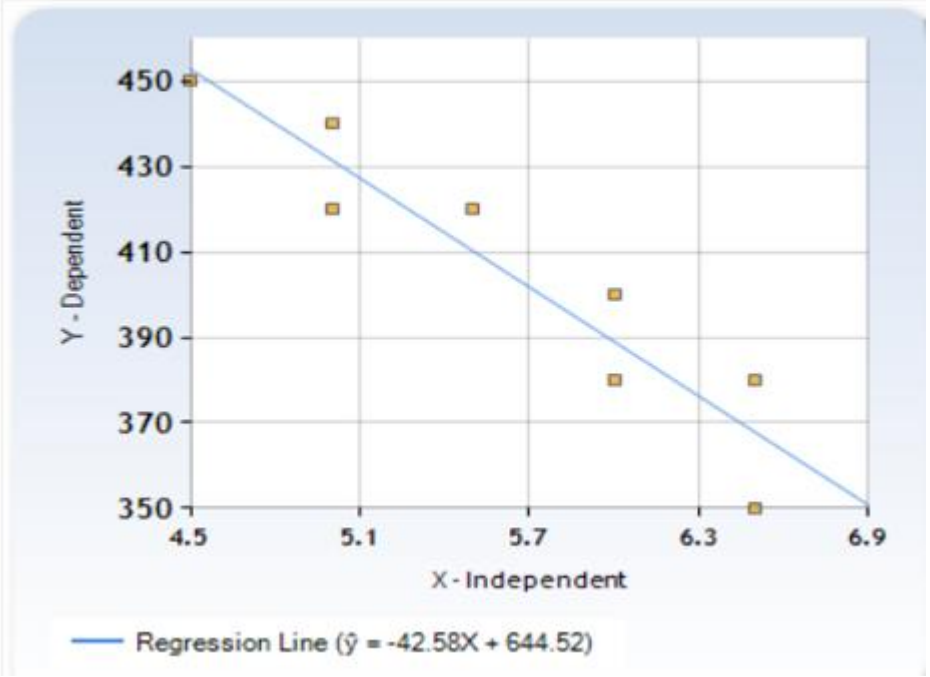
Use case II:

Estimate the effect of each independent variable (X) on the dependent variable (Y)

Effect of independent variable X on dependent variable Y

(a)What effect would you expect a $100 increase in price to have on sales?

A $100 increase in the rate will be expected to cause a 42.58 unit drop in LOC.



Use case III:

Calculate the correlation between the dependent variable and the independent variables.

Regression statistics table

Regression Output	Result	Explanation
Multiple R	??	R = square root of R^2
R Square	??	R^2
Adjusted R Square	??	Adjusted R^2 used if more than one x variable
Standard Error	??	This is the sample estimate of the standard deviation of the error
Observations	8	Number of observations used in the regression (n)

How to calculate R squares?

R squares is the percentage of the variance explain by the regression ($SS_{Regression}$) from the overall variance (SS_{Total}).

$$SS_{Regression} = \frac{[\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X})]^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2} = \frac{S_{xy}^2}{SS_x SS_y}$$

$X - M_X$	$Y - M_y$	$(X - M_X)^2$	$(X - M_X)(Y - M_y)$
-0.125	15	0.0156	-1.875
0.375	-25	0.1406	-9.375
0.875	-55	0.7656	-48.125
0.375	-5	0.1406	-1.875
-0.625	35	0.3906	-21.875
0.875	-25	0.7656	-21.875
-1.125	45	1.2656	-50.625
-0.625	15	0.3906	-9.375
		SS: 3.875	SP: -165

Sum of X = 45
Sum of Y = 3240

Mean X = 5.625
Mean Y = 405

Sum of squares (SS_X) = 3.875
Sum of products (SP) = -165

Calculate R square or coefficient of determination

$$SS_{Regression} = \frac{[\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X})\,]^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2\ \sum_{i=1}^{n}(Y_i-\bar{Y})^2} = \frac{S_{xy}^2}{SS_x SS_y} = R^2$$

Sum of X = 45
Sum of Y = 3240

Mean X = 5.625
Mean Y = 405

Sum of squares (SS_X) = 3.875
Sum of products (SP) = -165

Y - M_y
15
-25
-55
-5
35
-25
45
15

????

R^2 = ??

Multiple R = ??

0.8781

R square and Multiple R Interpretation

R^2 = 0.8025 means that 80.25% of the variation of y_i around ybar (its mean) is explained by the regressor X.

Multiple R: Correlation between y and y-hat is 0.8958

Note: The above R square and multiple R is not calculated from given LOC-Rate data.

Understanding Adjusted R Square

Adjusted R-squared

Essentially, the adjusted R-squared looks at whether additional input variables are contributing to the model. Consider an example using data collected by a pizza owner, as shown below:

Temperature (Celsius) X1	Price of Dough X2	Price of Pizza Y1
21	1	5
15	3	12
16	6	15
21	8	19
27	12	24
24	15	27
21	17	29
23	21	31
21	26	36

Assume the pizza owner runs two regressions:

Regression 1: Price of Dough (input variable), Price of Pizza (output variable)

Regression 1 yields an R-squared of 0.9557 and an adjusted R-squared of 0.9493.

Regression 2: Temperature (input variable 1), Price of Dough (input variable 2), Price of Pizza (output variable)

Regression 2 yields an R-squared of 0.9573 and an adjusted R-squared of 0.9431.

Why, the adjusted R-squared comes in?

Although temperature should not exert any predictive power on the price of a pizza, the R-squared increased from 0.9557 (Regression 1) to 0.9573 (Regression 2).

A person may believe that Regression 2 carries higher predictive power since the R-squared is higher. Even though the input variable of temperature is useless in predicting the price of a pizza, it increased the R-squared.

Which model prediction is more reliable ?

The adjusted R-squared looks at whether additional input variables are contributing to the model.

The adjusted R-squared in Regression 1(Dough) was 0.9493 compared to the adjusted R-squared in Regression 2(Temperature) of 0.9431.

Therefore, the adjusted R-squared is able to identify that the input variable of temperature is not helpful in explaining the output variable (the price of a pizza).

In such a case, the adjusted R-squared would point the model creator to using Regression 1 rather than Regression 2.

Which model Should be Used ??

Which model should be used? Information regarding both models are provided below:

	Model 1	Model 2
Variables Used	X1, X2, X3, Y1	X1, X2, Y1
R-squared	0.5923	0.5612
Adjusted R-squared	0.4231	0.3512

Formula of Adjusted R Square

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

where:

N is the number of points in your data sample.

K is the number of independent regressors, i.e. the number of variables in your model, excluding the constant.

How Adjusted R square Penalizes You?

The adjusted R^2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.

The adjusted R^2 will penalize you for adding independent variables (K in the equation) that do not fit the model. Why? In regression analysis, it can be tempting to add more variables to the data as you think of them. Some of those variables will be significant, but you can't be sure that significance is just by chance. The adjusted R^2 will compensate for this by that penalizing you for those extra variables.

While **values are usually positive,** they can be **negative** as well. This could happen if your R^2 is zero; After the adjustment, the value can dip below zero. This usually indicates that your model is a poor fit for your data. Other problems with your model can also cause sub-zero values, such as not putting a constant term in your model.

Use case IV:

Sample estimate of the standard deviation of the error.
Standard Error/ RMSE

Standard Error/ RMSE

There is a version of the formula for the standard error in terms of Pearson's correlation:

$$SE = \sigma = \frac{\sqrt{(1 - R^2) * SS_y}}{\sqrt{N - 2}}$$

Sum of X = 45
Sum of Y = 3240

Mean X = 5.625
Mean Y = 405

Sum of squares (SS_X) = 3.875
Sum of products (SP) = -165

Y - M_y
15
-25
-55
-5
35
-25
45
15
????

SE = ??

Todays Homework: Fill in the Table

Regression Output	Result	Explanation
Multiple R	??	R = square root of R^2
R Square	??	R^2
Adjusted R Square	??	Adjusted R^2 used if more than one x variable
Standard Error	??	This is the sample estimate of the standard deviation of the error
Observations	8	Number of observations used in the regression (n)

Todays Homework: Try Quiz

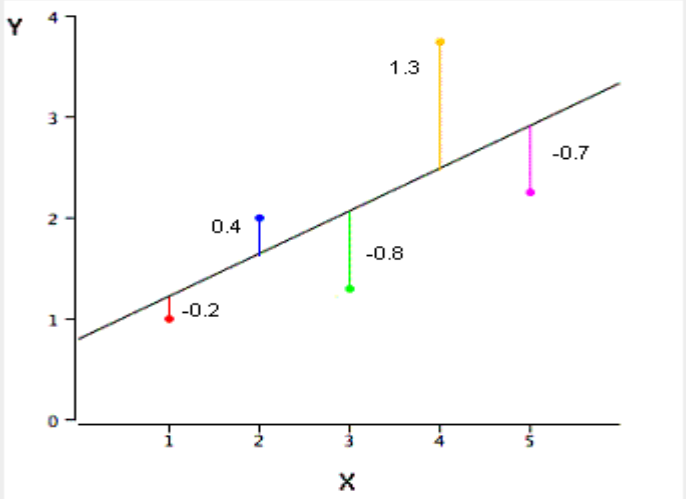
Question 1 out of 4.
In a regression line, the _____ the standard error of the estimate is, the more accurate the predictions are.

- ○ larger
- ○ smaller
- ○ The standard error of the estimate is not related to the accuracy of the predictions.

Question 2 out of 4.
Linear regression was used to predict Y from X in a certain population. In this population, SSY is 50, the correlation between X and Y is .5, and N is 100. What is the standard error of the estimate?

[ ]

Question 3 out of 4.
The graph below represents a regression line predicting Y from X. This graph shows the error of prediction for each of the actual Y values. Use this information to compute the standard error of the estimate in this sample.



ANOVA

Analysis of Variance Table