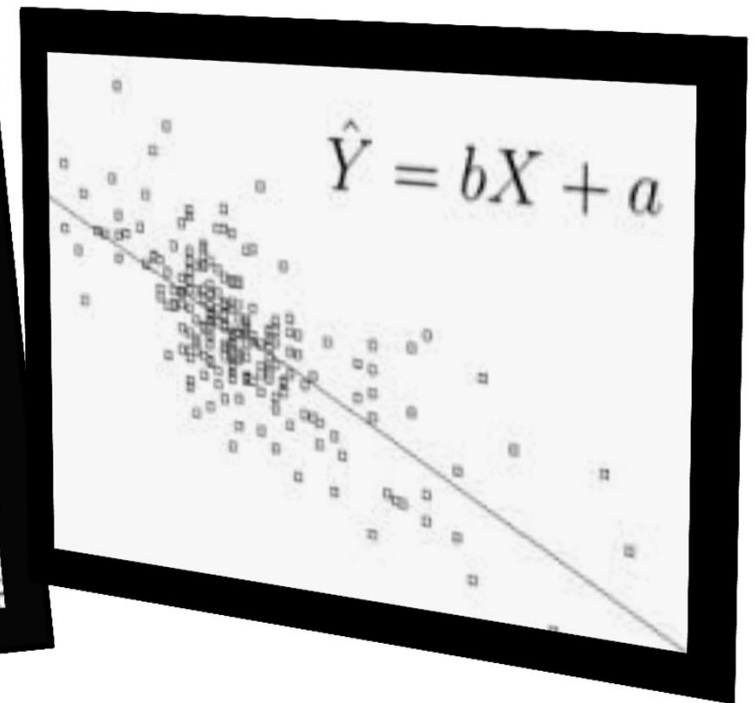
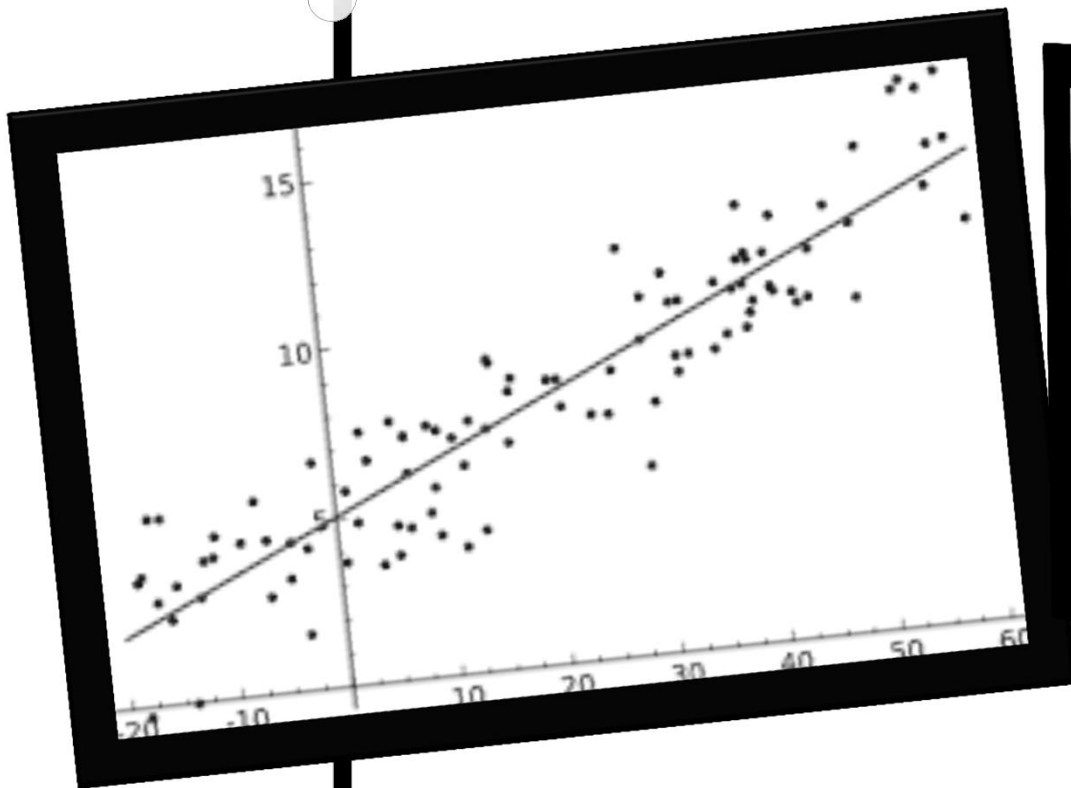


DATA MINING

LINEAR REGRESSION





Introduction

- Regression is a method to find the correlation between variables.
- Regression analysis is more accurate since the level of variable changes to other variables can be defined.
- The estimation of dependent variable value will be more accurate.



Introduction

- In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable Y and one or more explanatory variables (or independent variables) denoted X .
- The case of one explanatory variable (independent variable) is called simple linear regression. For more than one explanatory variable (independent variable), the process is called multiple linear regression.



Function of Linear Regression

- To find the average value of estimation and the value of dependent variable based on its independent variables.
- To test the dependency characteristics of hypothesis.
- To estimate the average value of independent variable based on the value of independent variable beyond the sample.



Data Analytics

1. Modeling the Regression Formula
2. Prediction
3. Determination Coefficient
4. Estimation Standard Error
5. Coefficient Standard Error
6. F Test
7. T Test
8. Summary



Criteria of Hypothesis Acceptance

- One-Tailed Test
 - A one-tailed test is used if only deviations in one direction are considered possible.
- Two-Tailed Test
 - A two-tailed test is used if deviations of the estimated parameter in either direction from some benchmark value are considered theoretically possible.



Example of Hypothesis Acceptance

One-Tailed Test

- H_o : There is no positive / negative influence of variable X against variable Y .
- H_a : There is positive / negative influence of variable X against variable Y .

Two-Tailed Test

- H_o : There is no influence of variable X against variable Y .
- H_a : There is influence of variable X against variable Y .
- H_o , accepted if $b \leq 0$, $t \text{ test} \leq t \text{ table}$
- H_a , accepted if $b > 0$, $t \text{ test} > t \text{ table}$



Simple Linear Regression Model

The formula of Linear Regression Y against X

$$Y = a + bX$$

Where:

Y = dependent variable

X = independent variable

a = intercept / constant

b = regression coefficient / slope

Calculating the Coefficient a and b

- Matrix Approach:

$$\begin{pmatrix} n & \Sigma X \\ \Sigma X & \Sigma X^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \Sigma Y \\ \Sigma XY \end{pmatrix}$$

$$a = \frac{\det A_1}{\det A} \quad b = \frac{\det A_2}{\det A}$$

$$A = \begin{pmatrix} n & \Sigma X \\ \Sigma X & \Sigma X^2 \end{pmatrix} \quad A_1 = \begin{pmatrix} \Sigma Y & \Sigma X \\ \Sigma XY & \Sigma X^2 \end{pmatrix} \quad A_2 = \begin{pmatrix} \Sigma n & \Sigma Y \\ \Sigma X & \Sigma XY \end{pmatrix}$$

$$\det A = (n)(\Sigma X^2) - (\Sigma X)(\Sigma X)$$

$$\det A_1 = (\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)$$

$$\det A_2 = (n)(\Sigma XY) - (\Sigma Y)(\Sigma X)$$

- Delivered:

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

$$b = \frac{(n)(\Sigma XY) - (\Sigma X)(\Sigma Y)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

$$\text{or} \quad a = \bar{Y} - b\bar{X}, \quad a = \frac{\Sigma Y - b(\Sigma X)}{n}$$



Determination Coefficient

- Determination Coefficient, denoted R^2 or r^2 and pronounced "R squared", is a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable.
- R^2 can be found:

$$R^2 = 1 - \frac{\sum (Y - Y_{pred})^2}{\sum (Y - Y_{rerata})^2}$$



Estimation Standard Error

- Standard Error is used to measure the error level of the regression model.
- The formula of Se:

$$Se = \sqrt{\frac{\sum (Y - Y_{pred})^2}{n - k}}$$

- Where:
 - n = the sample size
 - k = the number of variables (including independent and dependent variables)

Regression Coefficient Standard Error

- Regression Coefficient Standard Error (Sb) is used to measure the error level of the regression coefficient (b).
- The Sb value can be calculated:

$$Sb = \frac{Se}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}}$$



F-Test (*Fisher Distribution*)

- F-Test is used to test the model accuracy to identify the model that best fits the population from which the data were sampled:
 - H_0 : accepted if $F_{\text{test}} \leq \underline{F_{\text{table}}}$
 - H_a : accepted if $F_{\text{test}} > \underline{F_{\text{table}}}$
- F-Test can be calculated:

$$F = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)}$$

T-Test (*T Distribution*)

- T-Test is used to determine if two sets of data are significantly different from each other.
 - H_o : accepted if $t_{test} \leq t_{table}$
 - H_a : accepted if $t_{test} > t_{table}$
- The T-test value can be calculated:

$$T_{test} = \frac{bj}{Sbj}$$

- Where bj = regression coefficient j , and Sbj = regression coefficient standard error j



Example of Case Study

- A manager will conduct a research to find that there is an influence of promotion cost against the revenue in some companies in *WaterGold*. The research takes samples of 8 same companies which has done the promotion. The test significance level $\alpha = 5\%$.



Problem Solving

1. Research Topics

- The influence of promotion cost against the revenue in the company.

2. Research Question

- Is there positive influence of promotion cost against revenue in the company?

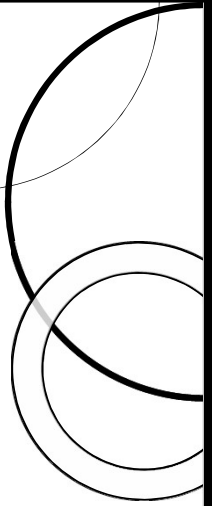
3. Hypothesis

- There is positive influence of promotion cost against revenue in the company.



Criteria of Hypothesis Acceptance

- H_o : There is no positive influence of promotion cost against revenue in the company.
- H_a : There is positive influence of promotion cost against revenue in the company.



Datasets (Train Data)

- Dataset of Promotion Cost and the Revenue

Promotion Cost (X)	Revenue (Y)
20	64
16	61
34	84
23	70
27	88
32	92
18	72
22	77



Calculation

No	Y	X	XY	X ²	Y ²
1	64	20	1280	400	4096
2	61	16	976	256	3721
3	84	34	2856	1156	7056
4	70	23	1610	529	4900
5	88	27	2376	729	7744
6	92	32	2944	1024	8464
7	72	18	1296	324	5184
8	77	22	1694	484	5929
Sum	608	192	15032	4902	47094
Average	76	24			



Regression Model

- Regression Coefficient (b)

$$b = \frac{[n \sum XY] - [(\sum X)(\sum Y)]}{[n \sum X^2 - (\sum X)^2]}$$

$$b = 1,497$$

- Constant (a)

$$a = \bar{Y} - b\bar{X}$$

$$a = 76 - (1,497 \times 24) = 40,1$$

- The Model:

$$Y = 40,1 + 1,497 X$$

Calculation

No	Y	X	XY	X ²	Y ²	Y _{pred}	(Y-Y _{pred}) ²	(Y-Y _{avg}) ²
1	64	20	1280	400	4096	70.014	36.163	144
2	61	16	976	256	3721	64.027	9.164	225
3	84	34	2856	1156	7056	90.966	48.525	64
4	70	23	1610	529	4900	74.503	20.281	36
5	88	27	2376	729	7744	80.490	56.403	144
6	92	32	2944	1024	8464	87.973	16.218	256
7	72	18	1296	324	5184	67.020	24.796	16
8	77	22	1694	484	5929	73.007	15.946	1
Sum	608	192	15032	4902	47094	608	227.497	886
Avg	76	24						



Determination Coefficient

- The Determination Coefficient (R^2):

$$R^2 = 1 - \frac{\sum (Y - Y_{pred})^2}{\sum (Y - Y_{average})^2}$$

- It is obtained:

$$R^2 = 1 - \frac{(227,497)}{(886)} = 0,743$$

- The Determination Coefficient(R^2) = 0,743, means the influence of promotion cost against the revenue is 74,3%. The remaining 25,7% caused by other factors which are not included in the model.

Estimation Standard Error

- Estimation Standard Error (Se):

$$Se = \sqrt{\frac{\sum (Y - Y_{pred})^2}{n - k}}$$

- Thus, the Se:

$$Se = \sqrt{\frac{(227,497)}{8 - 2}} = 6,1576$$

- Se = 6,1576, means that the limits of the prediction value may slip from the real data is at 6.1576

Regression Coefficient Standard Error

- The Sb can be calculated using formula:

$$Sb = \frac{Se}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}}$$

- Thus, the Sb :

$$Sb_1 = \frac{6,1576}{\sqrt{(4902) - \frac{(192)^2}{8}}} = 0,359$$



F-Test (*Fisher Distribution*)

- The F_{test} can be calculated using formula:

$$F_{\text{test}} = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)}$$

- Thus, the F test:

$$F = \frac{0,743 / (2 - 1)}{1 - 0,743 / (8 - 2)} = 17,367$$

- Since the F test (17,367) > F table (5,99) means that the regression model is good for use (*good of fit*).

T-Test (*T Distribution*)

- The t_{test} can be found using the formula:

$$T_{test} = \frac{bj}{Sbj}$$

- Thus, the t_{test} :

$$t_{test} = \frac{1,497}{0,359} = 4,167$$

- Since the t-test (4,167) > t table (1,943) thus the hypothesis H_a is accepted, means that there is positive influence of promotion cost against the revenue.



Summary

- Conclusion

There is positive influence of promotion cost against revenue in a company.

- Implication

It is recommended for companies to continue their promotion to increase the revenue.

It's easy, isn't it?





Exercise

- Find the simple linear regression model from the given dataset. Determine if the tree diameter has positive influence for the height.

No	Diameter (X)	Height (Y)
1	8	35
2	9	49
3	7	27
4	6	33
5	13	60
6	7	21
7	11	45
8	12	51