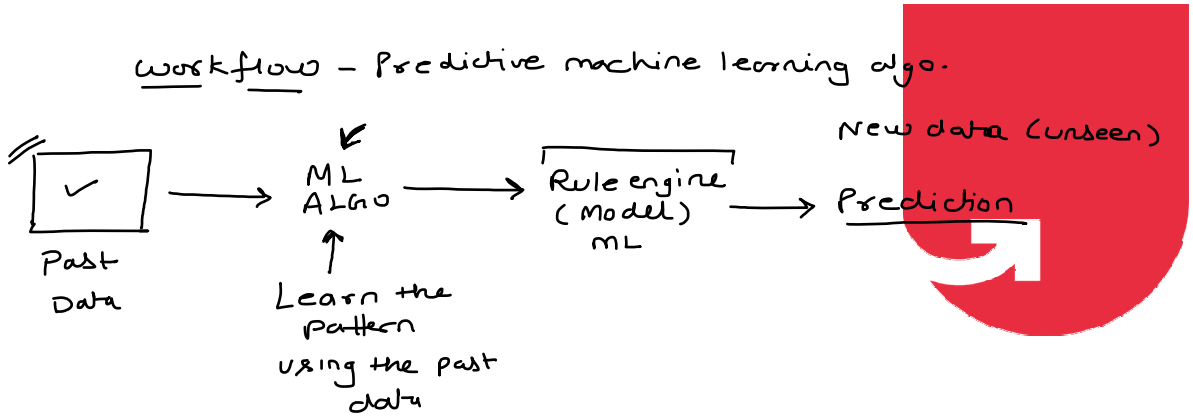
upGrad

#LifeKoKaroLift



Data Science: Linear Regression Doubts and Concepts

→ Predicts a continuous variable.

Past data

# of bed rooms	# of bathrooms	Squar area	Price

The attribute we are trying to predict is continuous in nature.

- → Predict sales based on marketing budget and other factors
- → Predict demand based on marketing strategy
- → Predict price of the house based various attribute about the data

-∞ to +∞

ML algo (Linear regression) → Rule

Price ~ f(# of bed, # of bath, Squar area)

2 - bed
1 - bath
500 sq feet

→ use this rule with new data to make prediction

→ Predicted price.

Course : Data Science

Lecture On : Linear Regression

Instructor : Sumit Shukla

upGrad

upGrad

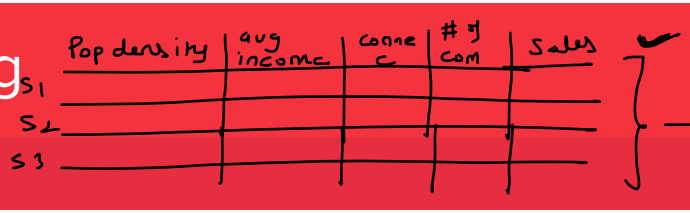
Today's Agenda

1. Regression Line
2. Best Fit Line
3. Strength of Simple Linear Regression
4. Need of Multiple Linear Regression
5. QNA

upGrad

#LifeKoKaroLift

Introduction to SLR



Introduction to Linear Regression

Choose that location which generate maximum sales

The factors that influences your target variable are independent variables

I am the CEO of a hypermarket chain and I want to open new store which should give me the best sales . I am hiring you as a Data Scientist to help me figure out a location where to open the new store

What factor will you be considering to open a new store?

what are the factors that are related to the location and will impact my sales at a particular store?

(independent variables)

- → Population density
- → avg income
- → connectivity
- → avg age
- → # of competitors
- → Parking
- → education level
- → standlone v/s shopping complex

- The hypermarket chain has more than 5000 stores across the world
- It is upstream hypermarket store catering to high end products
- There are more than 100 locations he needs to choose from

[Sales = f(pop.density, avg income, connect., avg age ....)]

dependent variable

Target

Sales = β1 (pop.density) + β2 (avg income) + β3 (connectiv) − β4 (# of competitor) + ...

Sales

weights / coefficients

Equation of straight line.

Introduction to Linear Regression

What could impact sales ?

- Population Density in the area
- Disposable Income
- Demographics of the region
- Parking size of the location
- No of other grocery stores in around (3km)
- Credit card usage
- Internet penetration/usage
- Average no of cars/household
- Avg family size/household
- No of working people/household

Sales = function (X1, X2, X3, X4,X5,X6……)

Sales = 10X1 + 20X2 +0.5X3 + 8X4 +……

If the function is linear we call it linear regression

Now this was a example of Multiple Linear Regression, Let's stop this topic for now and think in terms of two variables only, Say X and y. This we will called as Simple Linear Regression.

Simple Linear Regression

We have historic data of Marketing Spend and Sales, Let's see that

Past data — Independent — dependent

Marketing Spend (In Lakhs)	Sales (In Crores)
127.4	10.5
364.4	21.4
150	10
128.7	9.6
285.9	17.4

Now using this data we want to predict what we need to keep the marketing budget to generate X amount of sales.

Simple Linear Regression

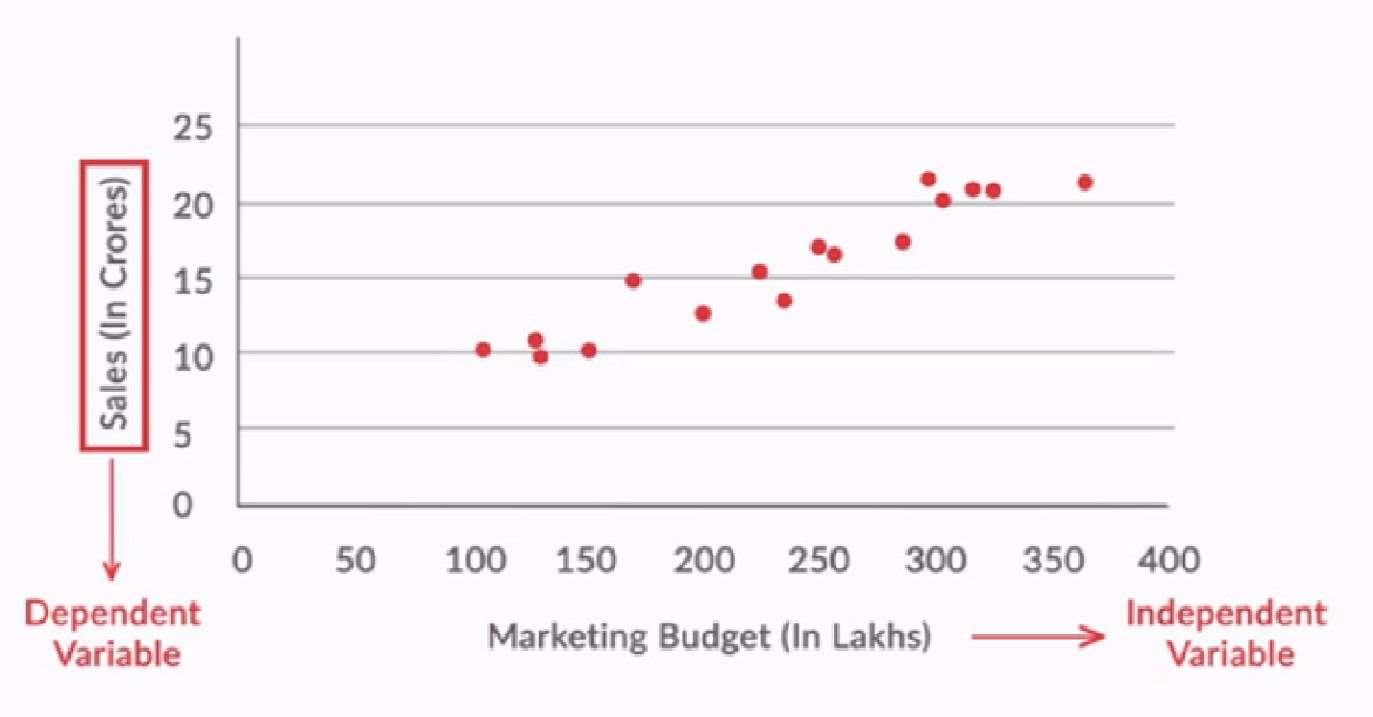
Sales $= \beta_1$ (marketing budget) $+ \beta_0$

$\Downarrow\Downarrow\Downarrow$

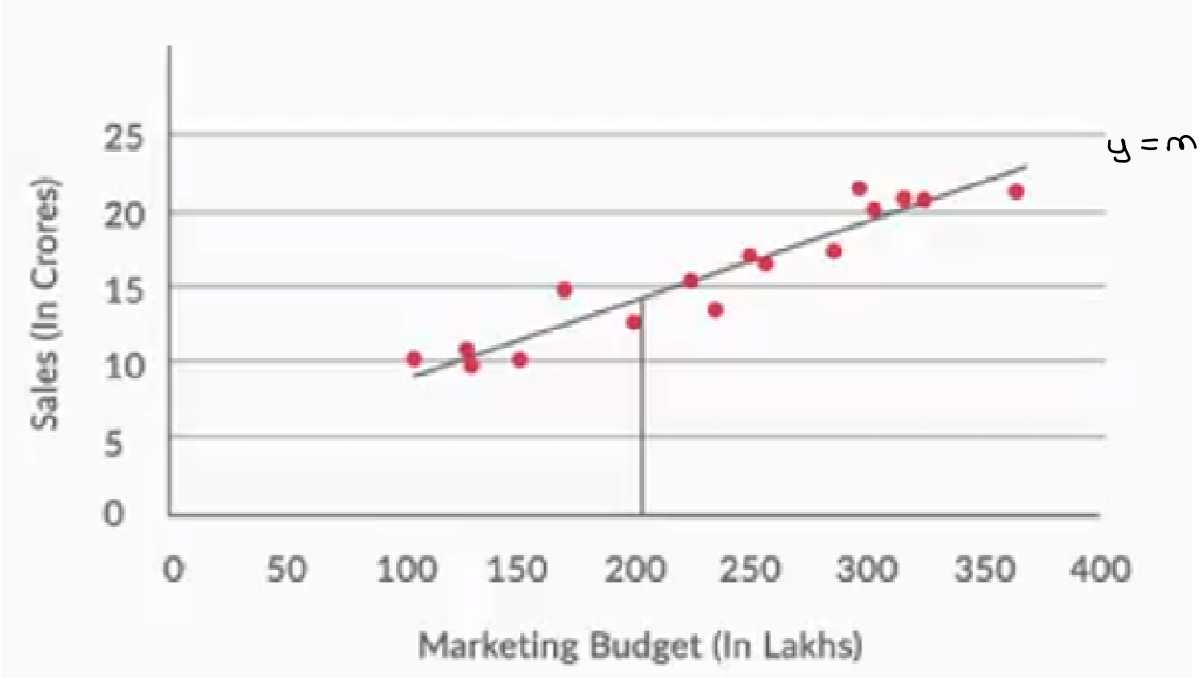
$y = mx + c$

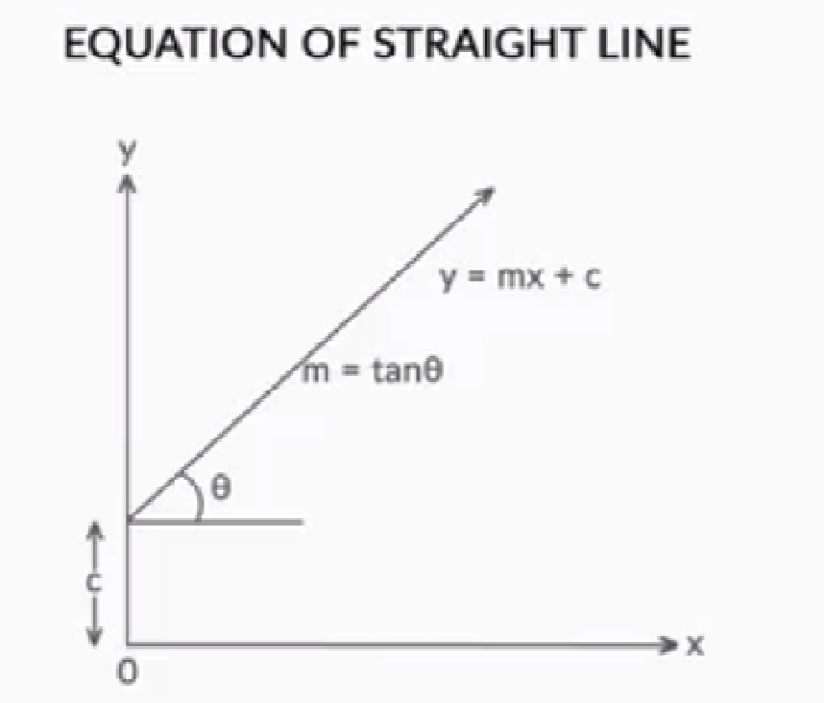
Let's first check if there is any relationship between Marketing Budget and Sales? If there is no relationship we would not put any money in our Marketing budget as this is not contributing to sales.

Simple Linear Regression



How to Make Prediction?





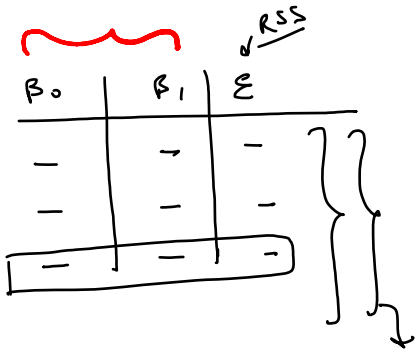
If there is zero marketing budget still we will observe C amount of sales.

← Is this a good model?

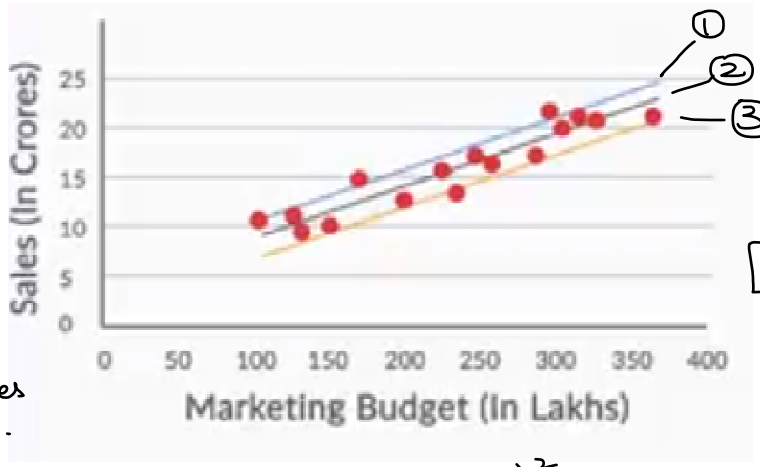
→ How to evaluate a regression model.

upGrad

Best Fit Line



OLS = ordinary least squares method.



$y = \beta_0 + \beta_1 x$ ①

$y = \beta_0' + \beta_1' x$ ②

$y = \beta_0'' + \beta_1'' x$ ③

[OLS out of all the possible lines (coefficients / weights), we will choose that particular line for which the error is minimum ≡ best fit line.]

$$RSS = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2$$

SLR

upGrad

$$\bar{y} = \frac{\sum y}{n}$$

x	y (actual)	$\hat{y}$ (Prediction)	$\varepsilon = (y_i - \hat{y})$ (Residual)	ε^2
1	3	3.5	-0.5	0.25
2	7	4	3	9
4	6	5	1	1
6	4	6	-2	4
8	8	7	1	1
	= 5.6	Total		15.25

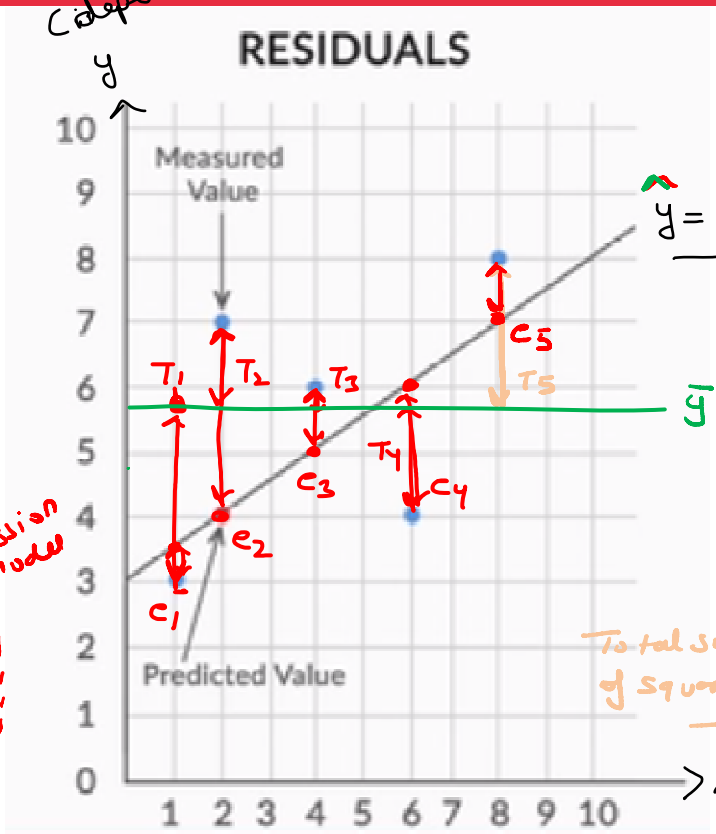
$$e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$$

$$\sum_{i=1}^{n} (y_i - \hat{y})^2$$

Error by Regression model

$$RSS = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

RSS → Residual sum of squares



$\hat{y} = \beta_0 + \beta_1 x$] Regression model

↓

model create using linear regression

$\bar{y}$ → avg-model. $(y_i - \bar{y})$

$$TSS = T_1^2 + T_2^2 + T_3^2 + T_4^2 + T_5^2$$

Total sum of squares

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

↓

Error by the avg model

SLR

Metric that can help us to understand if the regression model is good or bad.

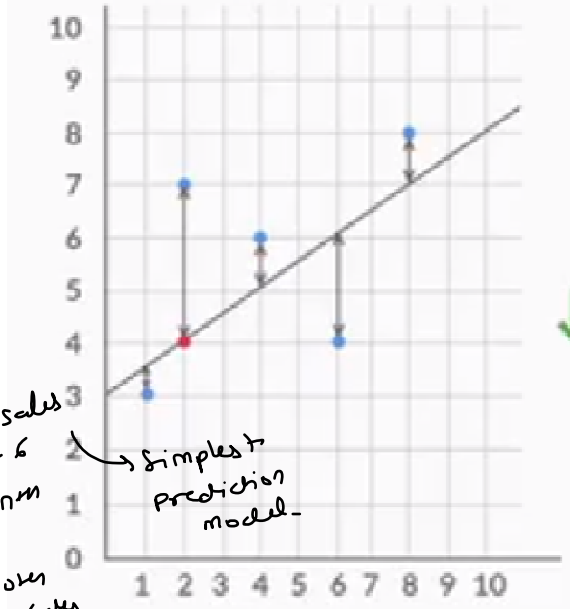
upGrad

Sales

1 —
2 —
3 —
4 —
5 —
6 —
7 —
8 —
9 —
10 —
11 —
12 —
1 —
2 —
3 —
1 —

avg sales last 6 months

new model approx sales



Best Fit Line

$Y = \beta_0 + \beta_1 X$

Intercept Slope

$e_i = Y_i - Y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\_\_ + e_n^2 = RSS$ (Residual Sum Of Squares)

$RSS = (Y_1-\beta_0-\beta_1 X_1)^2 + (Y_2-\beta_0-\beta_1 X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1 X_n)^2$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

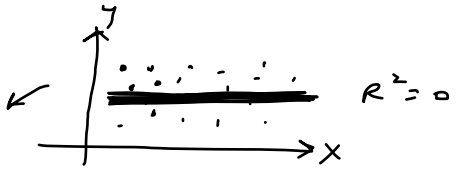
You — Father — Shormaji ka ladka.

reference

avg model

RSS = TSS $R^2 = 0$

RSS < TSS : improvement



RSS and TSS

Coefficient of determination

$R^2 = 1 - \frac{RSS}{TSS}$ → Error by regression model

→ error by avg-model.

80%.

[Your regression model is 80% better than the avg-model.]

Demo using: https://da-upgrad.shinyapps.io/rsquared/

$$TSS = (Y_1 - \bar{Y})^2 + \ldots + (Y_n - \bar{Y})^2$$

Or $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where

RSS - Residual sum of squares

TSS - Total sum of squares

TSS represent error of the very basic model that we can build without having any independent variable. So any model that we build should be better from this basic model.

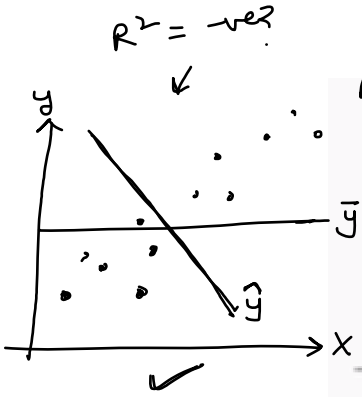
$TSS < RSS$

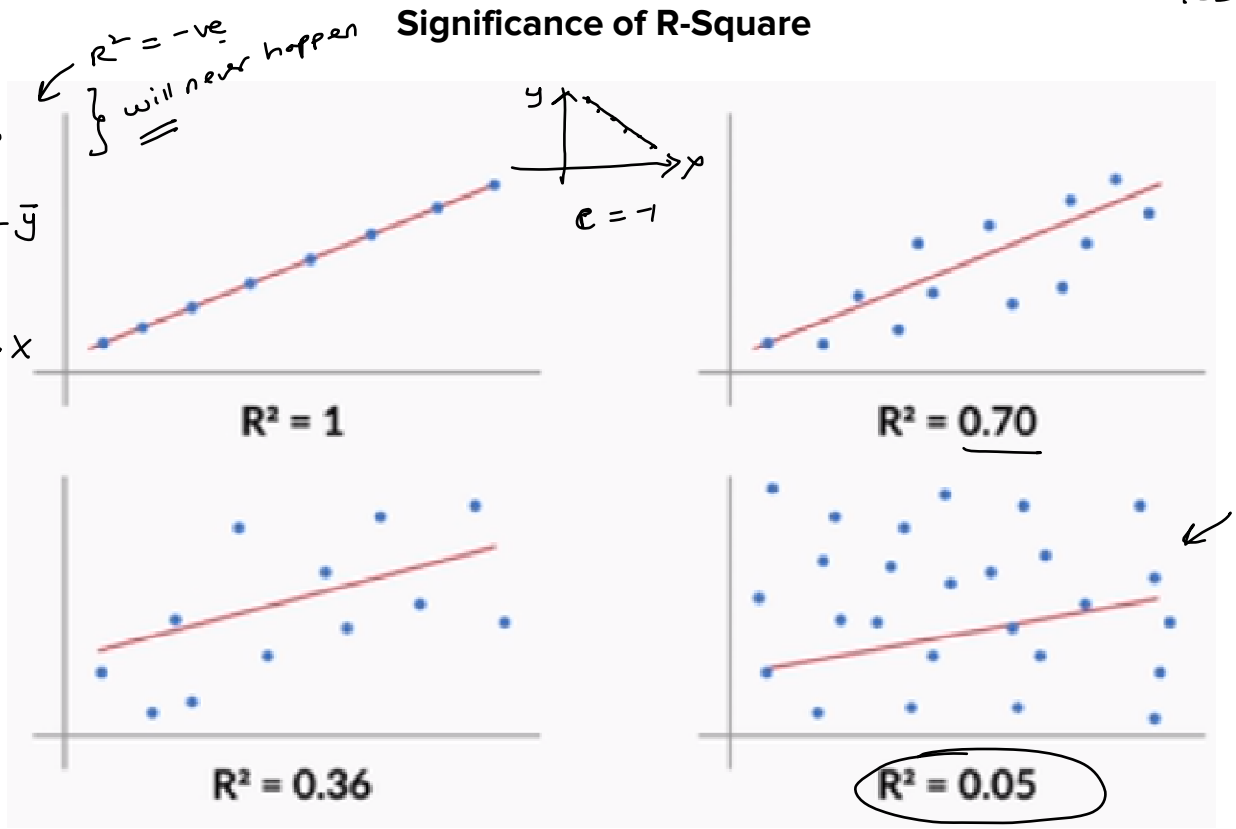
$R^2 = 1 - \frac{RSS}{TSS}$

$\downarrow R^2 \propto \frac{1}{TSS} \uparrow$

upGrad

Significance of R-Square





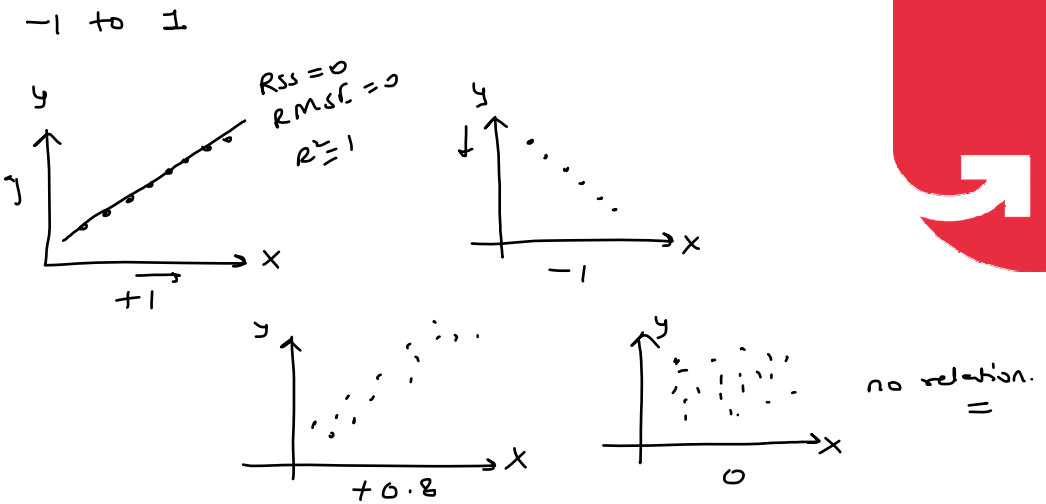
no relationship between x and y

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

The smaller the means squared error, the closer you are to finding the line of best fit.

upGrad

#LifeKoKaroLift



Let's Test our Knowledge

Let's Test

$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{2.75}{6.8} = 0.59 \approx 60\%$

upGrad

$\frac{\sum y_i}{n}$

$\Rightarrow$ my regression model is 60% better than the avg model.

X	Y_i	$\hat{y} = 1.5 + 1.5X$ Predicted $\hat{Y}$ Value	$(y_i - \hat{y})$ Error	$(y_i - \hat{y})^2$ Error Squared	Distance between Y values and their mean $(y_i - \bar{y})$	$(y_i - \bar{y})^2$ Mean distances squared
1	3	$1.5 + 1.5 \times 1 = 3$	$3 - 3 = 0$	0	$3 - 4.8 = -1.8$	3.24
2	4	$1.5 + 1.5 \times 2 = 4.5$	$4 - 4.5 = -0.5$	0.25	$4 - 4.8 = -0.8$	0.64
2	5	$1.5 + 1.5 \times 2 = 4.5$	$5 - 4.5 = 0.5$	0.25	$5 - 4.8 = 0.2$	0.04
3	6	$1.5 + 1.5 \times 3 = 6$	$6 - 6 = 0$	0	$6 - 4.8 = 1.2$	1.44
4	6	$1.5 + 1.5 \times 4 = 7.5$	$6 - 7.5 = -1.5$	2.25	$6 - 4.8 = 1.2$	1.44
Mean:	4.8	Sum:		RSS 2.75	Sum:	TSS 6.8

upGrad

#LifeKoKaroLift

Thank You!