upGrad

#LifeKoKaroLift

Data Science: Linear Regression Doubts and Concepts


upGrad
Course : Data Science
Lecture On : Linear Regression
Instructor : Sumit Shukla
2

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

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?

- 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

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

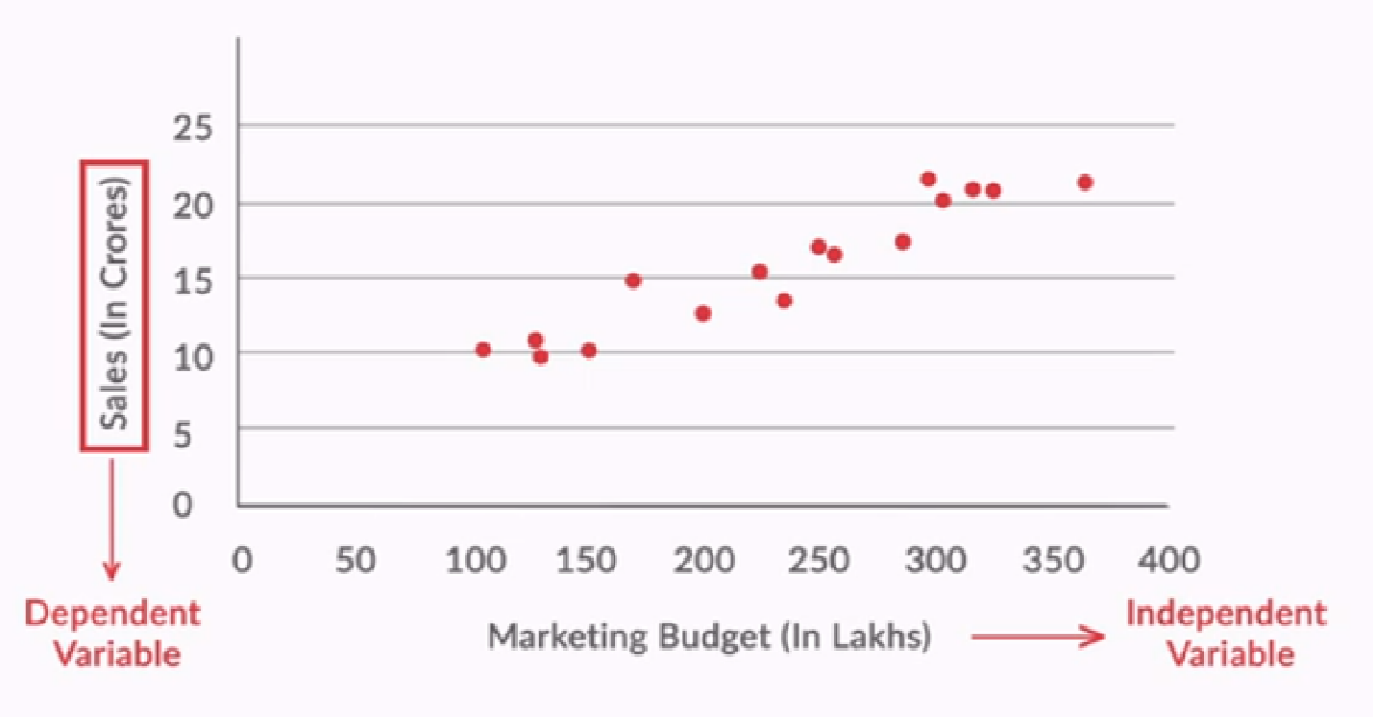
Marketing Spend (In Lakhs)	Sales (In Crores)
127.4	10.5
364.4	21.4
150	10
128.7	9.6
285.9	17.4

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

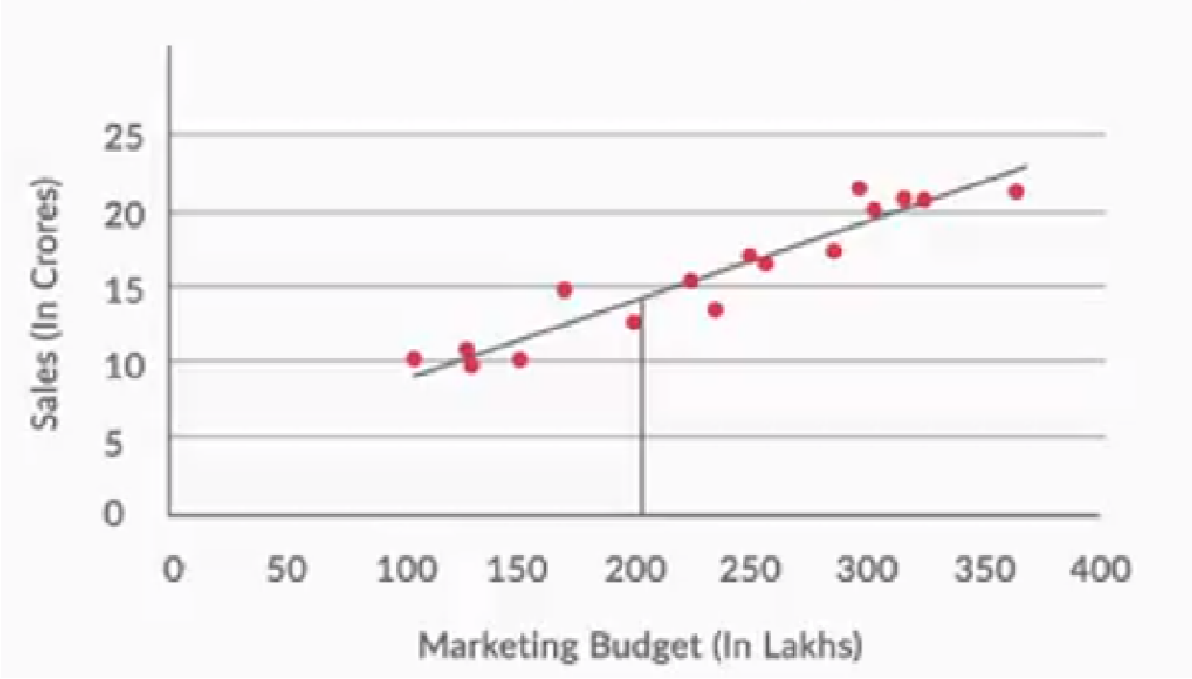
Simple Linear Regression

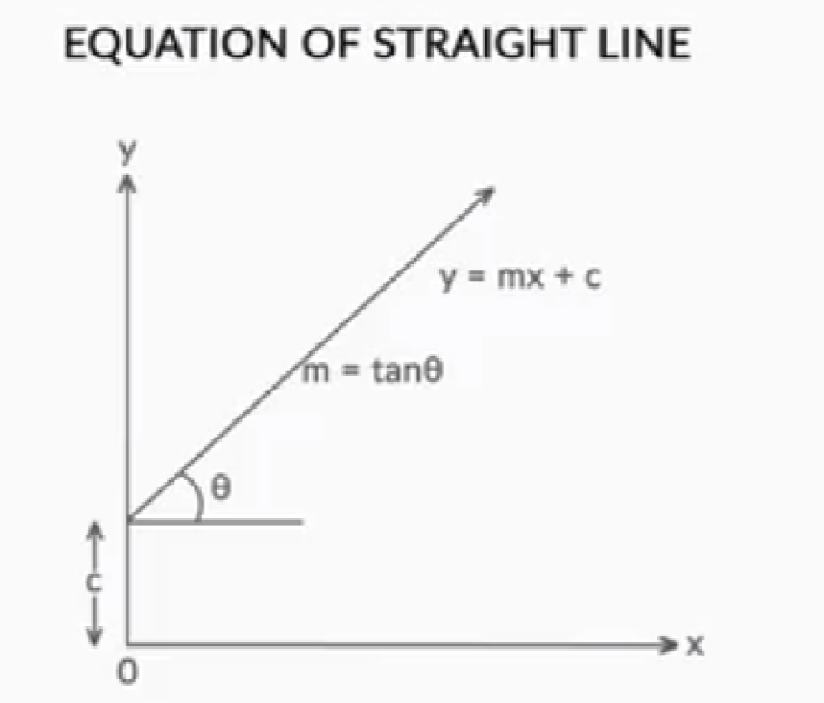
Let's first check if there is any relationship between Marketing Budget and Sales? If there is no relationship we would not put any money in our Marketing budget as this is not contributing to sales.

Simple Linear Regression



How to Make Prediction?

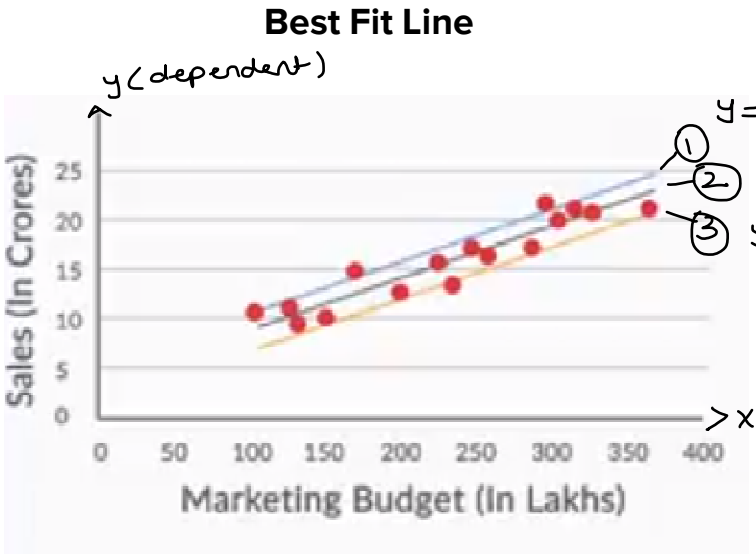




If there is zero marketing budget still we will observe C amount of sales.

SLR

Easy to handle while processing ← Smaller term



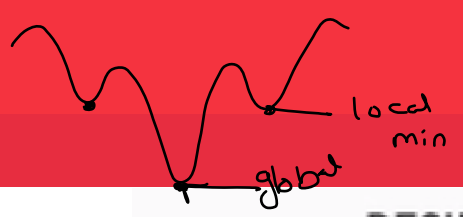
$y = \beta_0 + \beta_1 x$ ① → RSS1/n

$y = \beta_0' + \beta_1' x$ ② → RSS2/n

$y = \beta_0'' + \beta_1'' x$ ③ → RSS3/n

RSS2/n out of all the possible options of β_0 & β_1, we will choose that particular line for with your RSS is minimum.

SLR



$J(x) = x^2$

upGrad

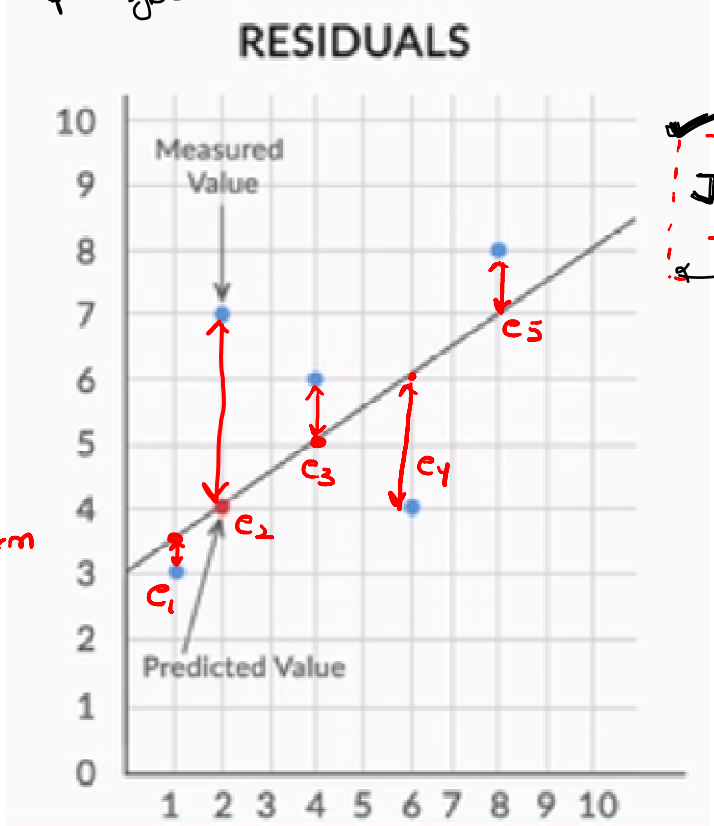
$RSS = e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$

$e = (y_i - \hat{y})$

$\hat{y} = \beta_0 + \beta_1 x$

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

find out that value of β_0, β_1 for which the RSS term is minimum

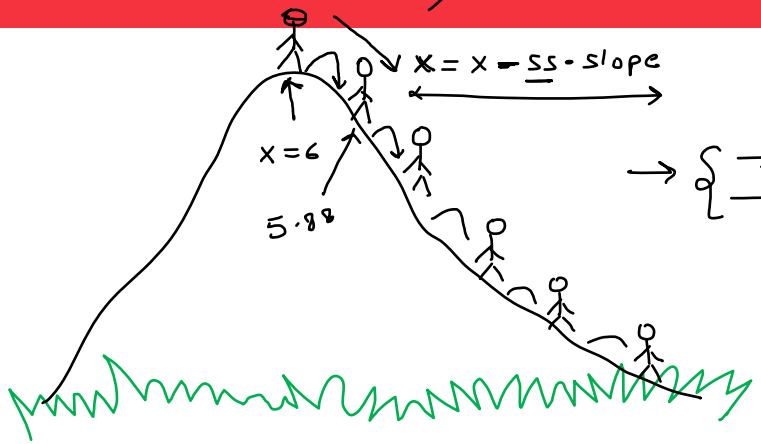


n = # of observation / rows in my data

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

Cost function for linear regression.

$x^4 = 4x^3 \frac{d}{dx} x^3$

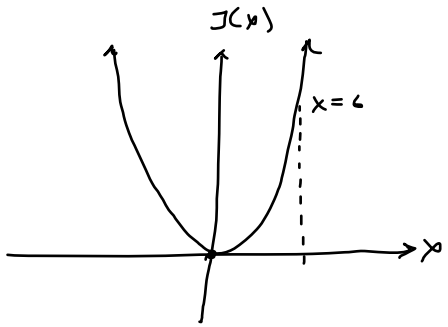


Gradient Descent → optimisation algo that helps you to find the minimum value of a given function.

→ Small step
→ In direction of slope.

$J(x) = x^2$

I need that value of x for which J(x) is minimum



Slope of any given function is the derivative of that function

$\frac{d}{dx} J(x) = \frac{d}{dx} x^2$

$= 2x^{2-1}$

$= 2x$

$\frac{d}{dx} x^n = n x^{n-1}$

$ss = 0.01$ slope $= 2x$

→ $x = 6$

→ next x = { current x − ss × slope }

$= 6 - 0.01 \times (2 \times 6)$

$= 6 - 0.12$

$= 5.88$

$5.88 - ss \times slope$

1000 iteration

$x = 0$

$$J(x) = x^2$$
$$= 2x^{n-1}$$
$$= 2x$$

$$J(x) = (x + 2x)^2$$
$$= 2(x+2x)^{2-1} \frac{d}{dx}(x+2x)$$
$$= 2(x+2x)\left\{\frac{d}{dx}x + \frac{d}{dx}2x\right\}$$
$$= 2(x+2x)(1+2)$$

$$\boxed{\frac{d}{dx}J(x) = 6(x+2x)}$$

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

find out that value of β_0 and β_1 for which $J(\beta_0, \beta_1)$ is minimum

Slope wrt β_0

$$\frac{\partial}{\partial \beta_0} J(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (y_i - (\beta_0 + \beta_1 x_i))^2$$

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

$$= \frac{1}{n} \sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_i)) \left\{ \frac{\partial}{\partial \beta_0} y_i - \frac{\partial}{\partial \beta_0} \beta_0 - \frac{\partial}{\partial \beta_0} \beta_1 x_i \right\}$$

$$= \frac{1}{n} \sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_i)) \{ 0 - 1 - 0 \}$$

$$= -\frac{2}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) = \boxed{-\frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{y})}$$

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

$$\frac{\partial}{\partial \beta_1} J(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} (y_i - (\beta_0 + \beta_1 x_i))^2$$

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

$$= \frac{2}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) \left\{ \frac{\partial}{\partial \beta_1} y_i - \frac{\partial}{\partial \beta_1} \beta_0 - \frac{\partial}{\partial \beta_1} \beta_1 x_i \right\}$$

$$= \frac{2}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) \{ 0 - 0 - x_i \}$$

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

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

$Corr(X_2, X_3)$ } Limited to only two series.
$Corr(X_2, X_4)$ }

$[X_2 \sim X_3, X_4, X_1]$

Multicollinearity: The presence of correlated independent features.

Rock band

- 2 Singers — male / female
- 1 Drum
- 1 keyboard
- 2 Guitar — same speed, same pitch, same tone

↓ delivering the same information

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad \} \; R^2 = 89\%$$

	using only independent variable	R^2	$vif = \frac{1}{1-R^2}$
1	$x_1 = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + \alpha_3 x_4$	10%	1.1
2	$x_2 = \alpha_0' + \alpha_1' x_1 + \alpha_2' x_3 + \alpha_3' x_4$	30%	1.42
3	$x_3 = \alpha_0'' + \alpha_1'' x_1 + \alpha_2'' x_2 + \alpha_3'' x_4$	50%	2
4	$x_4 = \alpha_0''' + \alpha_1''' x_1 + \alpha_2''' x_2 + \alpha_3''' x_3$	80%	5

[x_1, x_2 and x_3 together are able to explain 80% of x_4]

$vif \geq 5$] Higher multicolinearity

upGrad

#LifeKoKaroLift

Thank You!

upGrad

What is Multicollinearity?

Imagine you went to watch a rock band's concert. There are 2 singers, a drummer, a keyboard player, and 2 guitarists. You can easily differentiate between the voice of singers as one is male and other is female but you seem to have trouble telling who is playing better guitar.
Both guitarists are playing on the same tone, same pitch and at the same speed. If you could remove one of them then it wouldn't be a problem since both are almost same.
The benefit of removing one guitarist is cost-cutting and fewer members in the team. In machine learning, it is fewer features for training which leads to a less complex model.

Consider the simplest case where Y is regressed against X and Z and where X and Z are highly positively correlated. Then the effect of X on Y is hard to distinguish from the effect of Z on Y because any increase in X tends to be associated with an increase in Z.

Another way to look at this is to consider the equation. If we write $Y=b0+b1X+b2Z+e$, then the coefficient b1 is the increase in Y for every unit increase in X while holding Z constant. But in practice, it is often impossible to hold Z constant and the positive correlation between X and Z mean that a unit increase in X is usually accompanied by some increase in Z at the same time.

upGrad

What is VIF?

A variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

Variance inflation factors range from 1 upwards.

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.