Linear Regression's

We use ML for estimating relationships.

Problem statement:

Prepare a marketing plan for a Car Company

- Manufacturer of the car
- model of the car
- Engine Size of the car
- Horsepower of the car

← Making a market plan → Recommend which info of the car to highlight → To make the most of sales

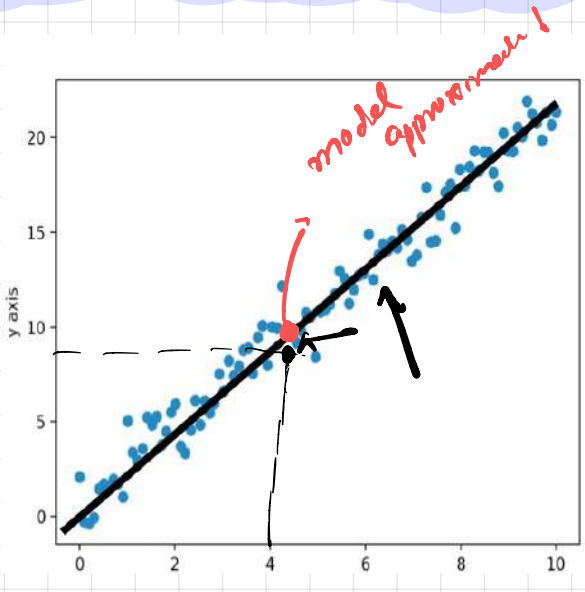
- Is there a relationship between info of the car and sales?
- How strong is the relationship between info of the car and sales?
- Which info contributes to sales?
- How accurately can we estimate the effect of each info on sales?
- How accurately can we predict the sales?
- Is the relationship linear?

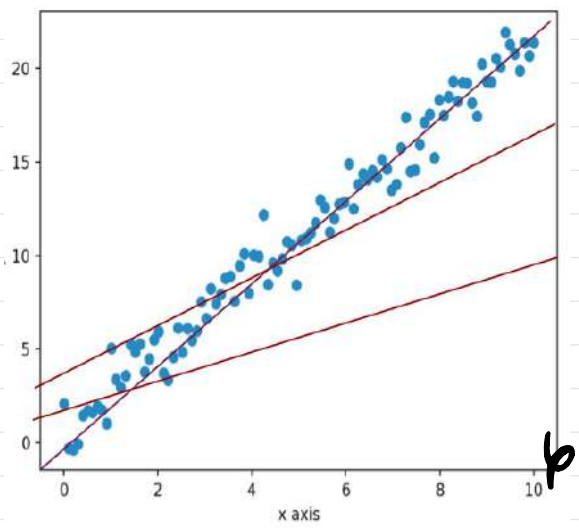
Linearity :-

→ data can be represented on the graph

→ Clear relationship b/w vars & that the graph will be straight line.

Geometrical understanding: Building a regression





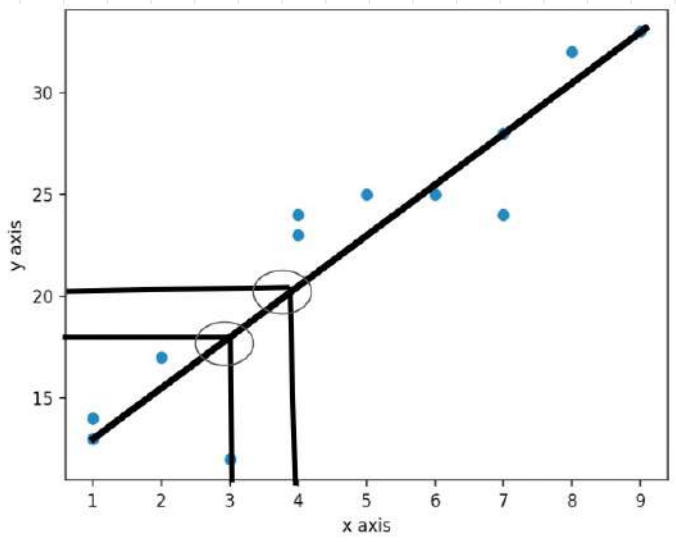
We try to find a straight line which best fits the data, so that we can make predictions.

Why find best fit line?

So let's say a person have 4 YOE, his predicted salary is 20k USD.

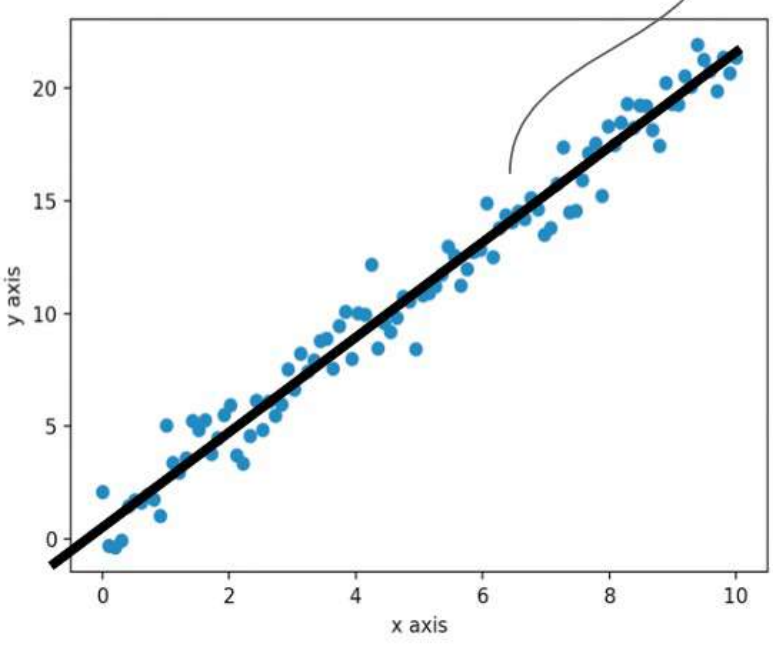
Basically we given an input x=4 and we found that y=20! We are mapping x to y using this best fit line.

Our data already tells a person have 3 yoe, his salary would be around 10K USD but model says his salary would be around 18K USD, that's why we want our resulting line to be as close as to the straight line.



So How do we find this best fit line?

How to get this?



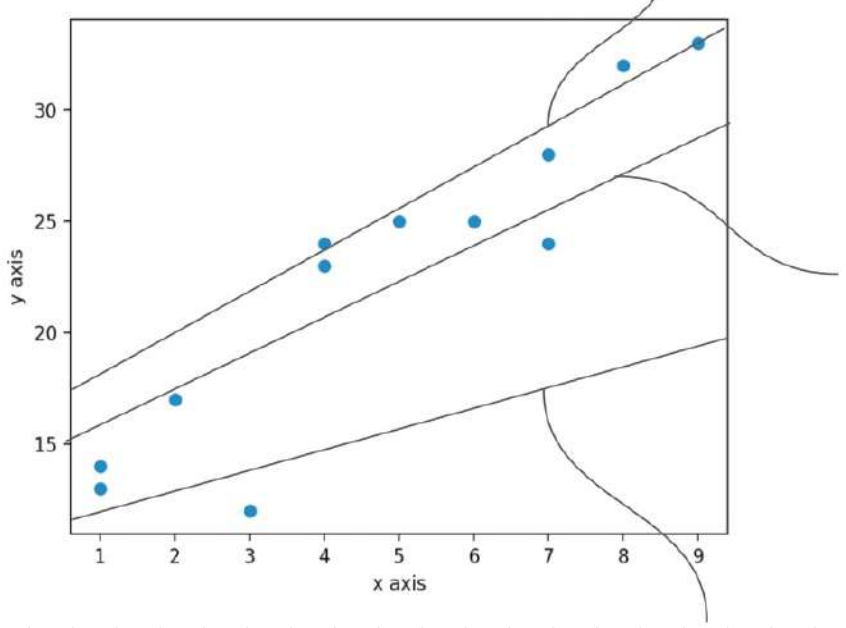
First solution: Use your drawing skills to draw a best fit line but there are major defects and we can't really make predictions from it.

Second solution: How this straight line formed mathematically?

$y = mx + b$ is a mathematical equation of a straight line,
m is slope of the line and
b is intercept.
You already have gone through in your earlier classes.

So, your straight line is dependent on m and b, if we're able to find best m slope and best intercept that fits the data, we got our new line :)

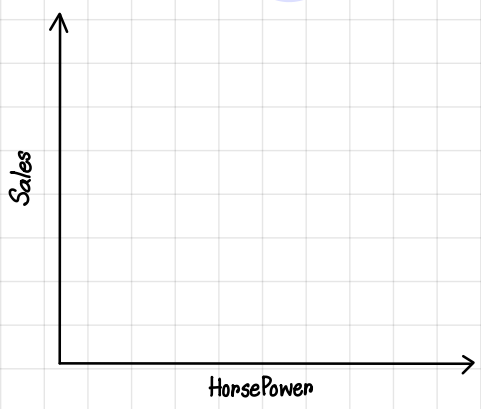
We try to find best "m" and "b".



This line has its own m and b, but does fits the data.

This line has its own m and b, but does not fits the data perfectly.

Dummy Example:



Horse Power	Sales
140	16
225	39
210	8
150	20
200	18

We need to find best fit line which fits the data, the resulting line is $y = mx + b$. So we have to find m and b.

$f(x) = mx + b$, we need to build a function that maps "x" Horsepower to "y" sales.

Notice: here we are multiplying m with x, so we can say that there is some weight m by which we are multiplying with x.

$f(\text{Horse Power}) = m(\text{Horse Power}) + y\text{-intercept}$

slope — where it intersects on y-axis

We also call this as weights or parameters.

$F(\text{Horse Power}) = m(\text{Horse Power}) + \text{intercept}$,

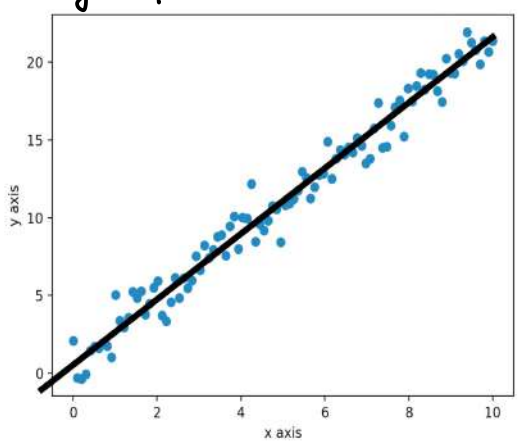
Let's assume m=3 and intercept = 4 (dummy) so we take our function as $f(x) = 3(x) + 4$, now we simply put x and then we get our prediction which is the sales.

How do we find the best m & b?

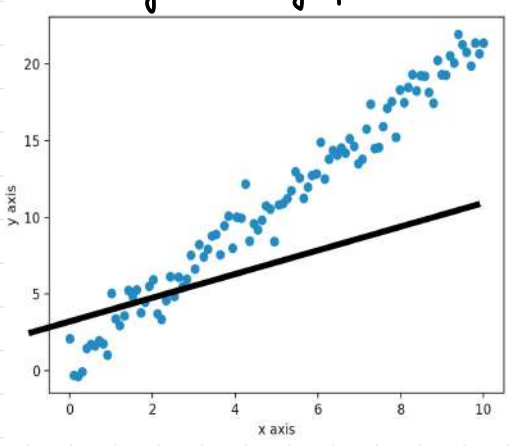
- One possible way to hit and trial different m and b values and see what works well. but It will be very hectic, Isn't it? and even impossible to do by hand.
- We have an optimization algorithm which is similar to our prev sol but it is done by bit of mathematical process!

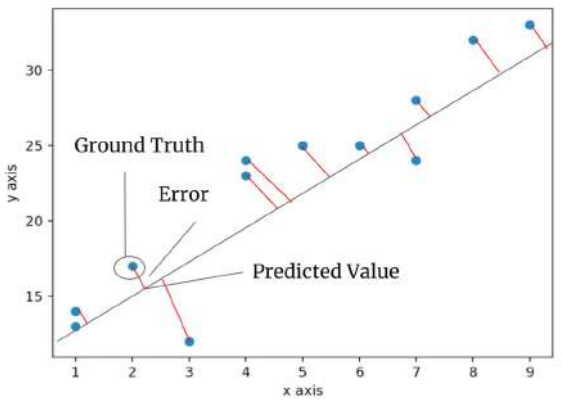
How to see what works well?

By Graphical

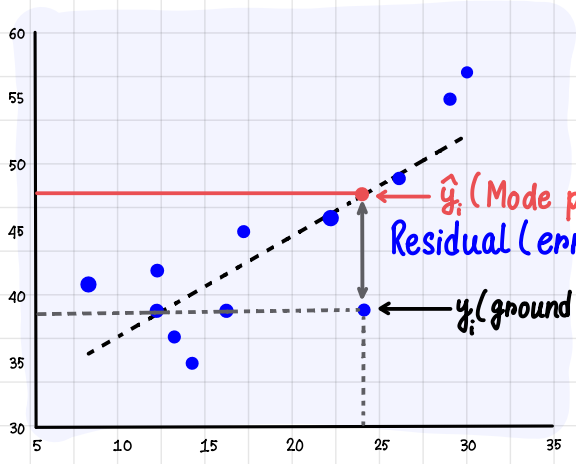


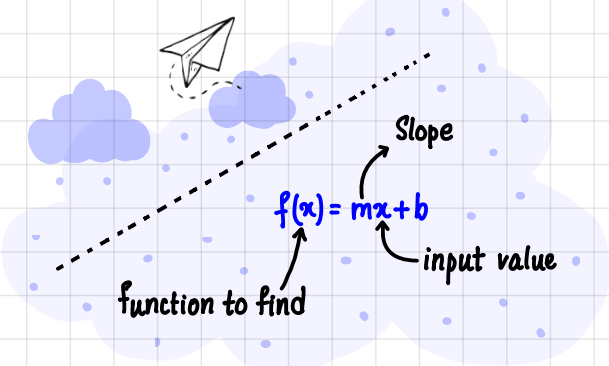
Just generalize graphical to numerical





Numerical Representation





For $x = 24$, approximated value or model predicted value is 47, but the ground truth $y = 38$. So the error term or error between y_i & $\hat{y}_i$ is $(\hat{y}_i - y_i) = (47 - 38) = 9$. So the error term is 9. Higher your error ie, bad your approximate is!

Similarly we can take out the e_i's for every data points out of here in the data

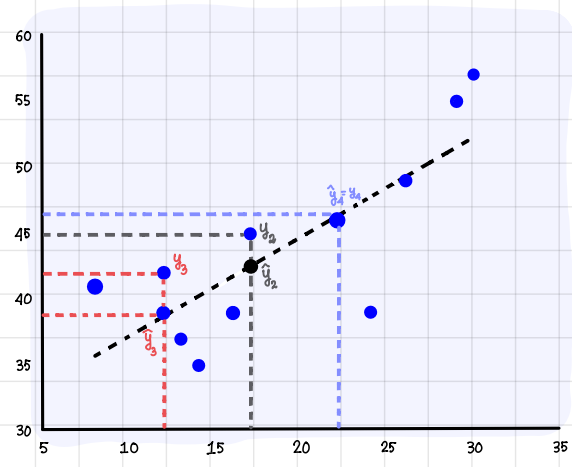
$(\hat{y}_2 - y_2) = (44 - 45) = -1$

$(\hat{y}_3 - y_3) = (39 - 43) = -4$

$(\hat{y}_4 - y_4) = (43 - 43) = 0$

⋮ for all data plots

$$(\hat{y}_i - y_i) = (n_1 - n_2) = e$$



Issues

a. we have m no of error terms.
b. some of our error terms is in "−"

Improving our baseline

squaring & absolute value

We have negative sign, we can get rid of it by:

- By squaring the no. $(\hat{y}_i - y_i)^2$
- By taking out the absolute $|\hat{y}_i - y_i|$

We have 'm' no of error terms, we can take average of m terms:

$$\frac{e_1 + e_2 + e_3 + \ldots + e_m}{m}$$

We can also take out average in 'Σ' format

$$= \frac{1}{m}\sum_{i=0}^{m} e_i \; ; \quad \frac{1}{m}\sum_{i=0}^{m} (\hat{y}_i - y_i)^2$$

adding sq to remove "−" sign; approx value; for averaging

Now, our error formula becomes:

$$\frac{1}{m}\sum_{i=0}^{m} (\hat{y}_i - y_i)^2$$ → Cost function (Mean Squared Error)

formula convention

$$J(m,b) = \frac{1}{2m}\sum_{i=0}^{m} (\hat{y}_i - y_i)^2$$

add 2 cause of "convention".

Why is it useful:

- It's way to evaluate the best fit line
- It's way to measure the performance how well your model perform on data.

We will take out the real reason of '2' later on

Worked example of MSE

Height(x)	Weight	h(x) = 0·8x + 9·2	errors
43	41	0·8 (43) + 9·2 = 43·6	41 - 43·6 = -2·6
44	45	0·8 (44) + 9·2 = 44·4	45 - 44·4 = 0·6
45	49	0·8 (45) + 9·2 = 45·2	49 - 45·2 = 3 8
46	47	0·8 (46) + 9·2 = 46	47 - 46 = 1
47	44	0·8 (47) + 9·2 = 46·8	44 - 46·8 = -2·8

Squaring all the error terms

$$= \frac{\text{Sum of error terms}}{m} = \frac{30·5}{5} = 6·08 \text{ (Ans)}$$

Bit about relationship

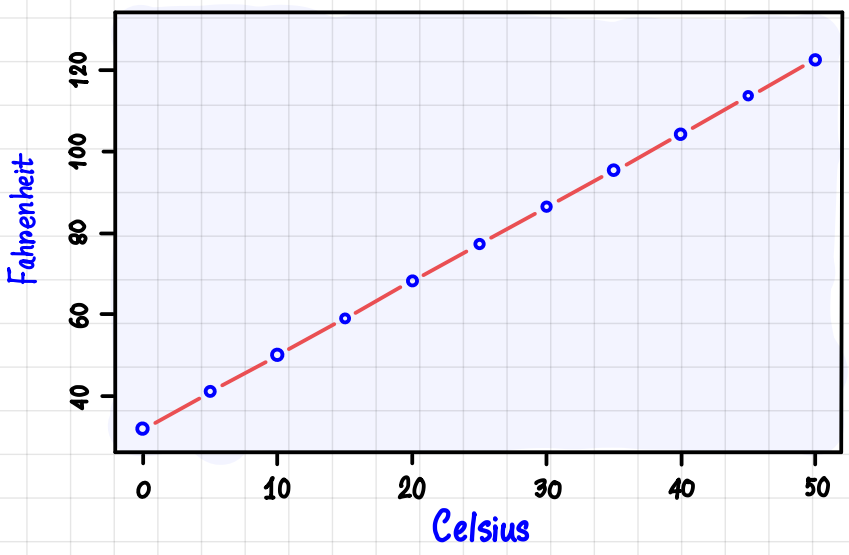
- Linear regression is statistical method that allows us to summarize data and study relationships between two Quantitative (continuous) variables / features.

Explanatory, independent or predictor value
- manufacturers
- model of the car
- enzine size
- horsepower

Response, predictor variable → Sales

types of relationships :

Deterministic reationship



$$F = \frac{9}{5}C + 32$$

↳ Relationship between degree 'F' and 'C'

If you know the value of C ; you know the value F exact.

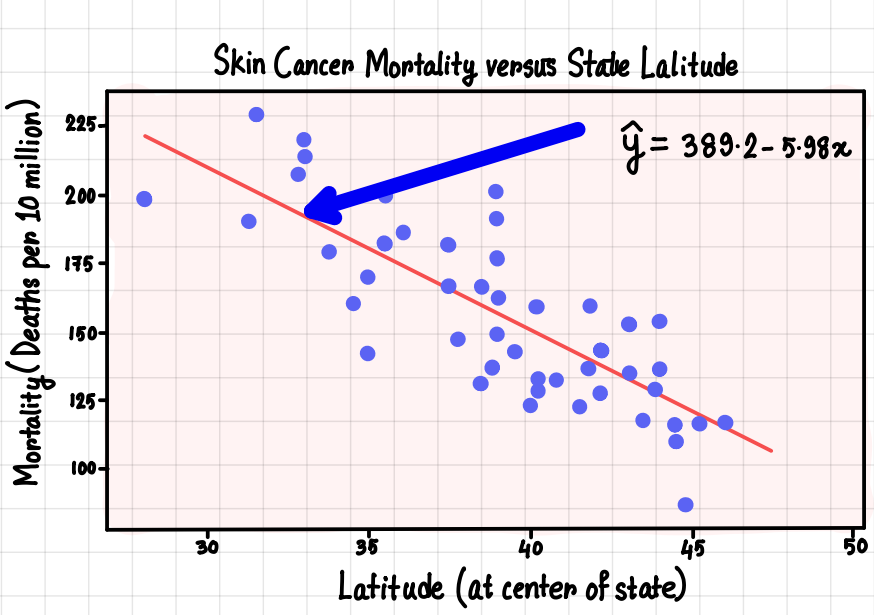
Others: $C = \pi \times \text{diameters}$

These are deterministic relationship, the equation exactly describes the relationship between two variables.

Find a relationship between skin cancer mortality & latitude

Higher altitudes of other northern u.s., the less exposed you'd be to the harmful rays of the sun, therefore less risk.

But the relationship isn't perfect, we don't get exact value.
- ⟶ It shows some trend.
- ⟶ It also shows scatter relationship.



Statistical relationship

We are interested in finding statistical relationship.

Estimating the coefficient

Given: $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (x_m, y_m)$

We try to find the best fit line based on given data which best fits the data.

Why is the best fitting line?

→ Line which best fits the data have 'n errors & they are as small as possible in some overall sense.

We have a way to achieve it by least squares criterion.

The form best fit line will take: $h(x_i) = mx + b$

We can evaluate our best fit line via MSE or least square criterion.

So, how to get good at errors?
↓
It depends on a straight line.
↓
Best fit line depends on two values m & b, the slope & an intercept.

Work on these → find good slope & intercept

We need to find m & b that minimize cost function.

Notational changes:

Min I(m,b) → Goal

$h(x) = mx + b$; $h(x) = \beta_1 x + \beta_0$

β_1, β_0 → parameters

β_1 → slope or feature weight

x → input value

β_0 → intercept or bias term

Given:

$y_i \approx \beta_1 x_i + \beta_0$ (for $i = 1, \ldots, n$)

How do we take out β_0 & β_1?

β_0 = intercept calculation

$\beta_0 = \bar{y} - b_1 \bar{x}$ ——— (1)

How do we take this?

We want to find β_1 & β_0 & so the resulting line is close as possible n points.

$$b_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{——— (2)}$$

WTH !!!

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

→ x_i = values of the x-variable in a sample.
→ $\bar{x}$ = mean of the value of the x-variable

Given: (1,1); (2,3); (4,5); (5,7)
$\bar{x} = 3$ & $\bar{y} = 4$

$$\beta_1 = \frac{(1-3)(1-4) + (2-3)(3-4) + (4-3)(5-4) + (5-3)(7-4)}{(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2} = \frac{(-2)(-3) + (-1)(-1) + (1)(1) + (2)(3)}{(-2)^2 + (-1)^2 + (+1)^2 + (2)^2} = \frac{6+1+1+6}{4+1+1+4} = \frac{14}{10} = 1.4$$

$\beta_0 = \bar{y} - b_1 \bar{x}$
$= 4 - 1.4 \times 3$
$= 4 - 4.2$
$= -0.2$

$h(x) = \beta_1 x + \beta_0$
$h(x) = 1.4 \times 1 + (-0.2)$
$= 1.4 - 0.2$
$= 1.2$

But what β_1 & β_0 tells?

Exploring β_1:

- If, $\beta_1 > 0$;
independent variable & response variable will have '+' relationship.

- If, $\beta_1 < 0$
independent variable & response variable will have '−' relationship.

x : predictor
y : dependent variable

x increases X↑
& ← ↓ lead ←
y increases Y↑

What is the '+' relationship?

Eg. House price prediction.

f(size) → price
↓ ↓
x → y

→ Increase in size of the house will lead to increase in the price.

$h(VOE) = \beta_1 VOE + \beta_0$
$h(0) = \beta_1 \times 0 + \beta_0$

Exploring β_0:

The average value of y when x is zero. (more on this in assignment)

Recap on what we have:

- hypothesis function: $h(x) = \beta_0 + \beta_1 x$
- MSE: $J(\beta_0, \beta_1) = \frac{1}{2} \sum_{i=1}^{m} (h(x) - y^{(i)})^2$
- Goal: Minimise MSE in order to find β_0 & β_1.
We will see an approach that is more common in finding best values of the parameters. β_0 & β_1

So, a simple question, "How to find $\beta_0 \times \beta_1$?

💡 Idea#1: Try out dweral values of β_0 & β_1 see if with that particular parameter, whether your cost function is decreasing on not.

- Initialize your $\beta_0=0$ & $\beta_1=0$; & you evaluate there, So, of c the cost function will be super high. Right?
- Now, you try out other value of β_0 & β_1 like 1.4 & 2.4 respectively, & see if cost function is decreasing.

And keep on doing this until & unless we get low error.

But this is a Big deal. → What to do?

We can keep on trying but we should know a strategy to keep on trying!! Right?

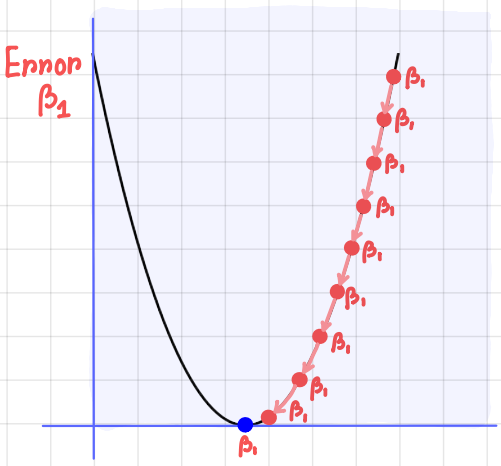
strategy ie what we want?

We want a system which automatically tries out the values every iteration in such a way that the iteration at n^{th} value 'parameter's value' is better than before iteration.

what is the strategy?

What is the plot of x^2?

So, $\frac{1}{2}\sum_{i=1}^{m}(\dots)^2$ → So the plot of the cost function will also look like a parabola.



This is what we want!!!

So, how this is possible?

Using the concept of Differentiation.

→ How much Y changes when X changes?

$\therefore \frac{\Delta Y}{\Delta X}$

→ How much $J(\beta)$ changes when β changes?

In a nutshell we are looking for this strategy only!!!

$\frac{dy}{dx} = \frac{dJ(\beta_0)}{d\beta_0}$; $\frac{dJ(\beta_1)}{d\beta_1}$

tells the same like how much $J(\beta_0)$ changes when β_0 changes → Same here as well.

$\beta_0 := \beta_0 - \alpha \frac{dJ(\beta_0)}{d\beta_0}$ → Lets expand the equation.

$\beta_0 := \beta_0 - \alpha \frac{dJ(\beta_0)}{d\beta_0}$

- → calculates the exact point using derivative at the current points.
- → substracts the value because we want to minimise the function
- → scale it by a factor of α

Gradient decent steps are:

- Initialize β_0 & β_1
- Calculate derivative/gradient at this point.
- Scaling it by a factor of α & minimise it by subtracting.
- Repeat 2 & 3 points.

For how much time we can repeat:

→ max. no. of iteration reached.

→ Step size is smaller than the tolerance.

$$\beta_1 := \beta_1 - \alpha \frac{dJ(\beta_1)}{d\beta_1}$$

→ doing the same for other parameters.

We can technically do for another parameters:

Repeat until {some conditions} {

$$\beta_i := \beta_i - \alpha \frac{dJ(\beta_i)}{d\beta_i} \quad (\text{for } i=0,1)$$

}

$\frac{dJ(\beta_i)}{d\beta_i} \longrightarrow \frac{\partial J(\beta_j)}{\partial \beta_j}$ (for $j=0,1$)

(for $i=0,1$) They both do same sort of work!!!

how indivual parameter effects MSE:

$$\frac{\partial J(\beta_j)}{\partial \beta_j} \; ; \; J(\beta) = \frac{1}{2}\sum_{i=1}^{m}\left(h(x)^{(i)} - y^{(i)}\right)^2$$

$$= 2.\frac{1}{2}\left(h(x)^{(i)} - y^{(i)}\right).\frac{\partial}{\partial \beta_j}\left(h(x)-y\right)$$

$$= (h(x)-y).x_j$$

$$\beta_j := \beta_j - \alpha \frac{(h(x)-y).x_j}{\frac{\partial J(\beta_j)}{\partial \beta_j}}$$ } LMS update rule

This is for single training example!!

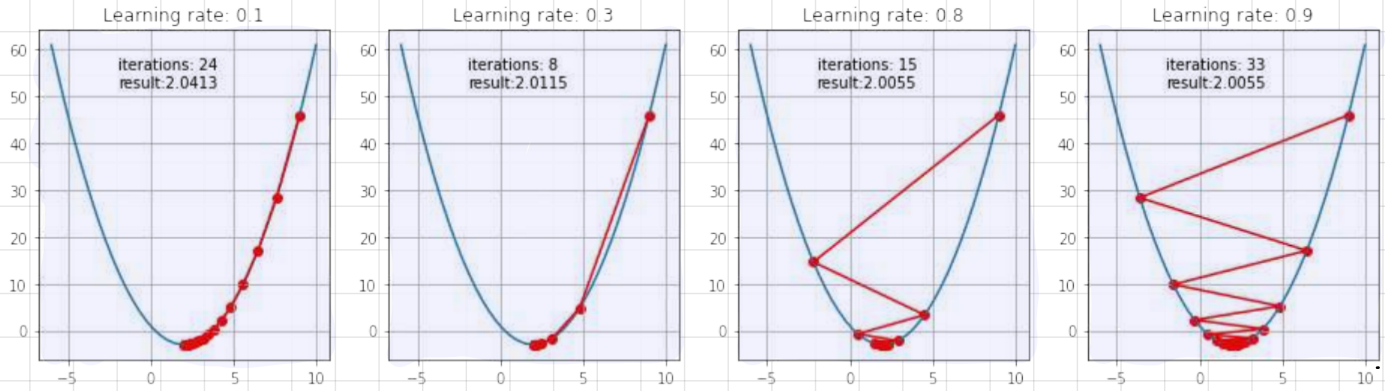
Repeat until some conditions {

$$\beta_i := \beta_i - \alpha \sum_{i=1}^{m}\left(h(x)^{(i)} - y^{(i)}\right)x_j^{(i)} \quad (\text{for every } j)$$

}

Story behind α:

"α" is a learning rate.



learning rate variations

Why does Gradient descent use derivative of the cost function?

- The first is the direction to move your Beta in.
- How big the step to take in?

If the slope is large we want to take a large step because we are far from the min. If the slope is small we want to take smaller step.

Hand Worked example

Look at google sheets provided in the reference resource.

Types of Requregression

Univariate linear regression:

You have only one input feature / variables based on that we map it to target variable.

$f(x) \longrightarrow y$

only one input feature

Multivariate linear regression:

So you want to build house price prediction:

- → size of the house (X_1)
- → No of Bedrooms (X_2)
- → No of fans (X_3)

} information (X_1, X_2, X_3) → Map Price of the house!!

how we can extend it to Multivariate linear regression?

If we have n features then the hypothesis becomes $h(x_1, x_2, \ldots, x_n)$

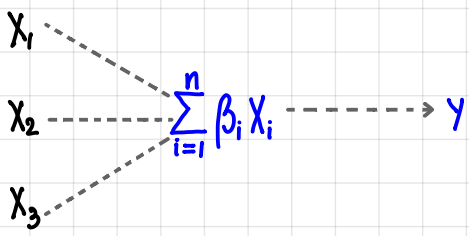
$h(X) \longrightarrow \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$

$[X_1, X_2, \ldots, X_n]$

* extend it to multiple features.

If we expand $\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \rightarrow \sum_{i=0}^{n} \beta_i X_i$

Vectorization
↓
convert a scalar vector to a vector program



$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \ldots & \ldots & \ldots & \\ \ldots & \ldots & \ldots & \\ X_{i1} & X_{i2} & \ldots & X_{ip} \\ \ldots & \ldots & \ldots & \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}$$

$n \times p$ (no of rows × no of columns)

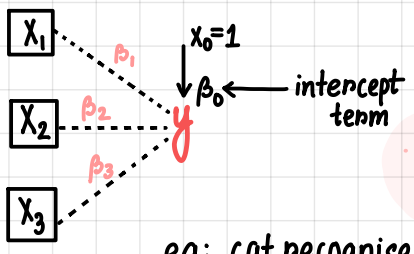
$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = X = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \ldots & \ldots & \ldots & \\ X_{i1} & X_{i2} & \ldots & X_{ip} \\ \ldots & \ldots & \ldots & \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}_{n \times p} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)}$$

$$X = \begin{bmatrix} (X^{(1)})^T \\ (X^{(2)})^T \\ \vdots \\ (X^{(m)})^T \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

$$J(B) = \frac{1}{2m} (X\beta - \vec{y})^T (X\beta - \vec{y})$$

↳ How we take this out?



eg: cat recognise
→ eyes (x_1) → β_1
→ ears (x_2) → β_2
→ mouth (x_3) → β_3

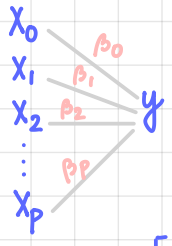
Generalise to p values?

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{bmatrix} = X \quad \to (p+1)\times 1$$

$$\vec{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$$

$$y \text{ or } h(x) = \underbrace{\beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{\text{hypothesis}}$$

$x_0 = 1$



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p$$

general regression equation

Matrix Distribution

$$X_{n\times p} = \begin{array}{c} \begin{matrix} \beta_1 & \beta_2 & . & \beta_i & . & . & \beta_p \end{matrix} \\ \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ x_{31} & x_{32} & \ldots & x_{3p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{i1} & x_{i2} & \ldots & x_{ip} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \end{array} \begin{matrix} \to \hat{y}_1 \\ \to \hat{y}_2 \\ \to \hat{y}_3 \\ \\ \to \hat{y}_i \\ \to \text{data matrix} \\ \to \hat{y}_n \end{matrix}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip}$$

$$\hat{y}_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \ldots + \beta_p x_{1p}$$
$$\hat{y}_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \ldots + \beta_p x_{2p}$$
$$\vdots$$
$$\hat{y}_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \ldots + \beta_p x_{np}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_n \end{bmatrix}_{n\times 1} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & & & & \vdots \\ 1 & x_{i1} & x_{i2} & \ldots & x_{ip} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}}_{n\times(p+1)} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_p \end{bmatrix}}_{(p+1)\times 1}$$

$n\times 1$

$$\hat{y} = X\beta$$

design matrix or data matrix $n\times(p+1)$

→ regression coefficient $(p+1)\times 1$

$$J(\beta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h(x^{(i)}) - y^{(i)}\right)^2$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & & & & \vdots \\ 1 & x_{i1} & x_{i2} & \ldots & x_{ip} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_p \end{bmatrix}$$

$n\times(p+1)$

let's break it down:

$$X = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ 1 & x_{31} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} \\ \beta_0 + \beta_1 x_{21} \\ \beta_0 + \beta_1 x_{31} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix}$$

$$\vec{e} = X\vec{\beta} - \vec{y} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} - y_1 \\ \beta_0 + \beta_1 x_{21} - y_2 \\ \beta_0 + \beta_1 x_{31} - y_3 \end{bmatrix}$$

$$\vec{e} \cdot (\vec{e})^2 = (X\vec{\beta} - \vec{y})^2 = \begin{bmatrix} (\beta_0 + \beta_1 x_{11} - y_1)^2 \\ (\beta_0 + \beta_1 x_{21} - y_2)^2 \\ (\beta_0 + \beta_1 x_{31} - y_3)^2 \end{bmatrix}$$

$$J(\beta) = \frac{1}{2m}(X\beta - \vec{y})^T (X\beta - \vec{y})$$

m n

Multiple liner Regression

Hypothesis:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & & & & \vdots \\ 1 & x_{i1} & x_{i2} & \ldots & x_{ip} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_p \end{bmatrix} \quad \vec{\beta}$$

$$h(X) = X\beta$$

input matrix → parameter vector!

Optimisation:

$$J(\beta) = \frac{1}{2m}(X\beta - \vec{y})^T (X\beta - \vec{y})$$

$\min J(\beta)$ ← goal to minimise cost function

Solve optimization problem:

a. Initialize the parameters β_i to "0".

b. Repeat {

$$\beta_{new} := \beta_{old} - \alpha\left[\nabla_\beta J\right]_{\beta_{old}}$$

c. $\nabla_\beta J = \left[\frac{\partial J}{\partial \beta_1}, \frac{\partial J}{\partial \beta_2}, \ldots, \frac{\partial J}{\partial \beta_j}\right]$

} algorithm to optimize the problem.

vectorized Gradient descent!! ← H.W.!!

Analytical Solution of Linear Regression:

What if we obtain β by simply plugging X and y into $\beta = (X^T X)^{-1} X^T y$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ i & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

$X^T X \longrightarrow$ is not inventible

- You may need to:
  - → remove extra features to ensure that $d \leq n$.

Gradient descent
- → Requires multiple iterations
- → Need to choose α.
- → Works well when n is large.
- → Can support incremental learning.

Closed form Solution
- → Non-iterative.
- → No need for α.
- → Slow if n is large
- → Computing $(X^T X)^{-1}$ is roughly $O(n^3)$.

Problem:

$$\vec{y} = \begin{bmatrix} 10 \\ 20 \\ 30 \\ 40 \\ 50 \end{bmatrix}_{5\times1} \quad X = \begin{matrix} X_0 & X \\ \end{matrix} \begin{bmatrix} 1 & 5 \\ 1 & 7 \\ 1 & 10 \\ 1 & 12 \\ 1 & 20 \end{bmatrix}_{5\times2} \quad \vec{\beta} = (X^T X)^{-1} X^T y$$

Step 1: $X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 5 & 7 & 10 & 12 & 20 \end{bmatrix}_{2\times5} \begin{bmatrix} 1 & 5 \\ 1 & 7 \\ 1 & 10 \\ 1 & 12 \\ 1 & 20 \end{bmatrix}_{5\times2} = \begin{bmatrix} 5 & 54 \\ 54 & 718 \end{bmatrix}$

$2\times2 \leftarrow$ matrix

Step 2: $(X^T X)^{-1} = \frac{1}{|X^T X|}$

Determinant →

$|X^T X| = \begin{vmatrix} 5 & 54 \\ 54 & 718 \end{vmatrix} = 5\times718 - 54\times54 = 674$

$\text{Adj}(X^T X) = \text{adj of} \begin{bmatrix} 5 & 54 \\ 54 & 718 \end{bmatrix} = \begin{bmatrix} 718 & -54 \\ -54 & 5 \end{bmatrix}$

So, $(X^T X)^{-1} = \frac{1}{674} \begin{bmatrix} 718 & -54 \\ -54 & 5 \end{bmatrix} = \begin{bmatrix} 1.07 & -0.0801 \\ -0.0801 & 0.007 \end{bmatrix}$

Step 3: Compute $X^T y$

$$X^T y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 5 & 7 & 10 & 12 & 20 \end{bmatrix}_{2\times5} \begin{bmatrix} 10 \\ 20 \\ 30 \\ 40 \\ 50 \end{bmatrix}_{5\times1} = \begin{bmatrix} 150 \\ 1970 \end{bmatrix}$$

Step 4: $\beta = (X^T X)^{-1} X^T y$

$$= \begin{bmatrix} 1.07 & -0.0301 \\ -0.0301 & 0.007 \end{bmatrix} \begin{bmatrix} 150 \\ 1970 \end{bmatrix} = \begin{bmatrix} 2070 \leftarrow \beta_0 \\ 10775 \leftarrow \beta_1 \end{bmatrix}$$

$y = \beta_0 + \beta_1 X_1$

$\hat{y} = 2.70 + 1.77(X) \longrightarrow$ regression equation.

Evaluating the Model

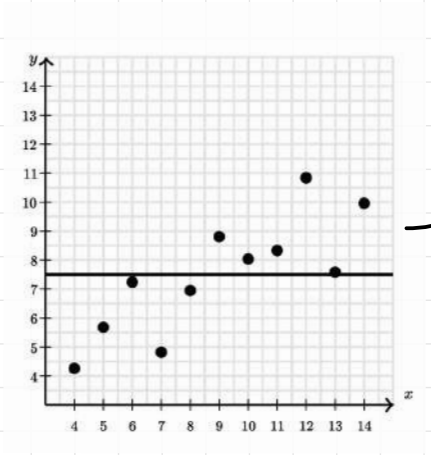
→ RSE : $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

↳ it measures the lack of fit of the model.

→ Higher the RSE bad the model is !!!

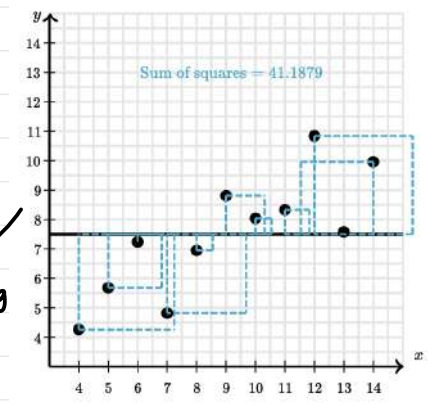
→ R^2 statistic :-

→ Say you wanna predict y without a corresponding X val. we can simply predict the average of all y values.

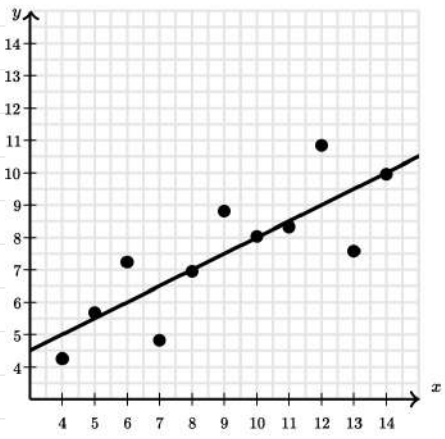


this not a best fit line.

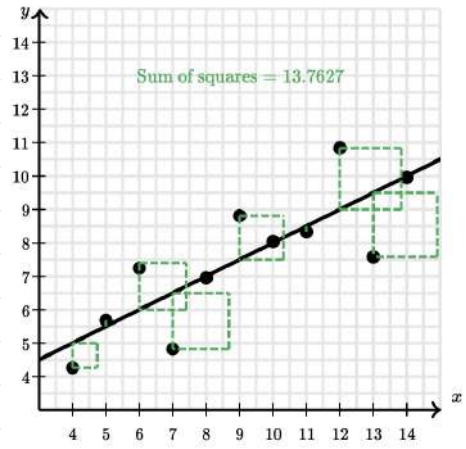
error = 41.1879



→ we can use linear regression to best fit the data.



error →



So using the L.R. we removed considerable amount of prediction error.

But how much?

A square measures how much pred. error we eliminated.

Total reduction = 41·1879 − 13·7627
= 27·4252 → $\frac{41·1879 - 13·7627}{41·1879} \approx 66·59\%$ → we can represent as %

→ it tells us what % of the pred. error in the y-var is eliminated when we use least squares regression on the x-var.
↳ Khan academy def.

$$R^2 = \frac{\text{Varience explained by the Model}}{\text{total variance}}$$

0% represents that a model doesn't explain any variations in the res. var. around its mean.
100% represents that a model explains all the variation in the response var. around its mean.

Limitations of this method

→ It doesn't tell whether you have chosen model is good or bad.
→ It will tell you whether the data spread are biased.
→ A high or low R-sq. isn't good or bad.

Multiple Linear Regression:

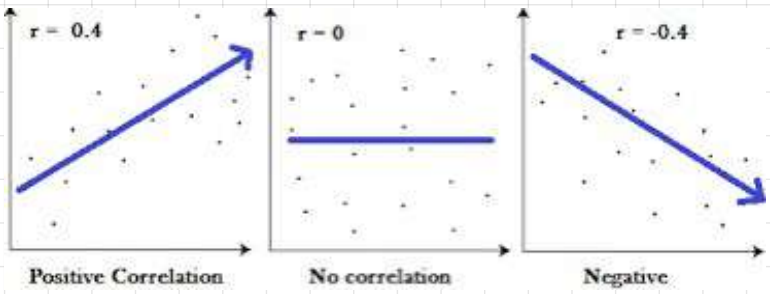
Q1. Is atleast one of the predictors $X_1, X_2, \ldots, X_p$ useful in pred. the response?
↳ is there any relationship between response & predicton?

- Correlation used to measure the strength of the relationship between 2 variables.

Correlation coefficient

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

→ $-1 \le r \le 1$, r takes the value between −1 & 1 inclusivity.
→ The sign indicates the direction of the lr. relationship - positively or negatively related.
→ A correlation of 0 means that there's no relationship.



eg:- look at eq: q-2 of lecture reading material.
eg:- look at eq: q-3

Returning back to our Question:
Is atleast one of the predictors $X_1, X_2, \ldots, X_p$ useful in pred. the response?
→ The null hypothesis states that the model with no independent vars fits the data.
→ The alternative hypothesis states that the model fits the data better than intercept only model.

Q. Steps of doing hypothetic tests with L.R models:
1. formulate 0 hypothesis about the relationship between the predictor and response variables.
e.g. hypothesize if there's a positive relationship between the predictor & response var.
→ That the predictor has no effect on response var.
* Specify a null hypothesis, which is a statement about the value of a population that is assumed to be true.
eg: The null hypothesis might be the slope of the regression line is zero.
↓
which means that there isn't any effect of predictor on response.

Hypothesis formulation for Test:

⟶ In case of l.R, the claim is made that there exist a relationship between response and predictor variables.
⟶ In the claim is represented using the non-zero coef. of predictor variables in the l.r. equations ⟶ alternative hypothesis.
⟶ Thus, the null hypothesis is set that there is no relationship between response & the pred var. hence the coef. =0 for each feature.

if $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
then null hypothesis
$y = 0 + 0 \times x_1 + 0 \times x_2$

for all predictor var., individual hypothesis testing is done to determine wether the relationship between response & the particular predictor var. about the significance.

Types of hypothesis tests which can be carried out in regression models

"Test for significance of regression"

t-test
"this test checks the significance of individual regression coefficient."
$h(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ → Overal Significance of regression coefficient.
→ can also be used to take out or test individual coefficients.

F test
"This test can be used to simultaneously check the significance of no. of regression equation.

Hypothesis formulation for f-Test:

⟶ Hypothesis test done around the claim that there is a linear regression model representing the response var. & all the predictor var.
⟶ The null hypothesis is that the l.R. model does not exist; which means that the value of all the coefficient is equal to 0.

Determining the test statistic:

→ T-statistics for testing hypotheses related to individual coefficients:

Test for significance by performing a t-test for the regression
Slope t
$H_0: \beta_1 = 0$
$H_A: \beta_1 \neq 0$

P-value:

⟶ A p-val. is a statistical measurement used to validate a hypotheses against observed data.
⟶ A p-val. measures the probability of obtaining the observed results assuming research on p-value by yourself!

Hypothesis Testing Readings

There are several ways to test the significance of a regression model, including the t-test, the F test, and the p-value.

The t-test is used to determine wether the regression coefficients for each predictor variable are significantly different from zero. A coefficient that is significantly different from zero indicates that there is a relationship between the predictor and response variables, and that the predictor can be used to predict the response.

The t-test is calculated as

$$t = \frac{b_i}{SE(b_i)}$$

where b_i is the coefficient for the ith predictor variable, and $SE(b_i)$ is the standard error of the coefficient.

The F-test is used to determine whether the overall regression model is significantly better than a model that only includes the intercept (ie. a model that does not use any predictor variables).

The F test is calculated as:

$$F = \frac{\left(\frac{MSR}{p}\right)}{\left(\frac{MSE}{n-p-1}\right)}$$

where MSR is the mean square regression, p is the number of predictor variables in the model, MSE is the mean square error; and n is the total number of observations.

The p-value is the probability that the observed relationship between the predictor and response variables could have occured by chance, given that there is no actual relationship. A small p-value (typically less than 0·05) indicates that the observed relationship is statistically significant, and that the regression model can be used to make predictions about the response variable. The p-value is calculated using the t-statistic and the degrees of freedom for the regression model.

In summary, the t-test, F-test and p-value are all used to evaluate the significance of a regression model, and to determine whether the predictor variables can be used accurately predict the response variable.

What is the sum of Squares?

In regression analysis, the sum of squares regression (SSR) is a measure of the variation in the response variable that is explained by the regression model. It is calculated as the sum of the squared differences between the predicted response values from the regression model and the mean of the observed response values.

The SSR is used to evaluate the improvement in the model's fit compared to using the mean of the response values as the predicted value for all observations. It is calculated as:

$$SSR = \sum(\hat{y}_i - \mu)^2$$

where $\hat{y}_i$ is the predicted response value for the ith observation, and μ is the mean of the observed response values.

The SSR can be used in conjunction with the sum of squares total (SST) and the sum of squares error (SSE) to evaluate the overall fit of the regression model. The SST is a measure of total variation in the response variable, and the SSE is a measure of the residual variation in the response variable that is not explained by the regression model. The SST and SSE are calculated as:

$$SST = \sum(y_i - \mu)^2$$
$$SSE = \sum(y_i - \hat{y}_i)^2$$

where y_i is the observed response value for the i-th observation and $\hat{y}_i$ is the predicted response value from the regression model.

In summary, the SSR is a measure of the variation in the response variable that is explained by the regression model, and is calculated as the sum of the squared differences between the predicted and mean response values. It is used to evaluate the improvement in the model's fit compared to using the mean of the response values as the predicted value for all observations.

Here is an example of calculating the sum of squares regression (SSR) is a simple linear regression model with two predictor variable (x_1 and x_2) and one response variable (y):

x_1	x_2	y
1	2	5
2	3	7
3	4	9
4	5	11

The regression equation for this model is:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

where b_0, b_1 and b_2 are the regression coefficients.

Assume that the estimated values for the coefficients are $b_0 = 1, b_1 = 2, b_2 = 3$

The predicted response values for each observation can be calculated using the regression equation:

$\hat{y}_1 = 1 + 2\times1 + 3\times2 = 9$,

$\hat{y}_2 = 1 + 2\times2 + 3\times3 = 15$,

$\hat{y}_3 = 1 + 2\times3 + 3\times4 = 21$,

$\hat{y}_4 = 1 + 2\times4 + 3\times5 = 27$

The mean of the observed response values is $(5+7+9+11)/4 = 8·5$

The SSR is calculated as the sum of the squared differences between the predicted response values and the mean of the observed response values:

$$SSR = (9-8·5)^2 + (15-8·5)^2 + (21-8·5)^2 + (27-8·5)^2 = 0·25 + 6·25 + 12·25 + 18·25 = 37$$

Therefore, the SSR for this regression model is 37. This indicates that the regression model explains a significant amount of the variation in the response variable compared to using the mean of the response values as the predicted value for all observations.

What is degrees of freedom associated with SSR?

In regression analysis, the sum of squares regression (SSR) is a measure of the variation in the response variable that is explained by the regression model. The degrees of freedom associated with the SSR are the number of observations in the data set minus the number of parameters in the regression model.

For example, if a data set contains 10 observations and a regression model has two parameters (the intercept and the slope), then the degrees of freedom for the SSR are 10-2=8. This means that the SSR is calculated using the variation in the response variable for the 8 remaining observations, after the effect of the regression model has been removed.

The degrees of freedom for the SSR are used in calculating the mean square error (MSE), which is the average of the squared differences between the observed and predicted values. The MSE is calculated as

$$MSE = \frac{SSR}{df}$$

where df is the degrees of freedom for the SSR.

In summary, the degrees of freedom for the SSR are the number of observations in the data set minus the number of parameters in the regression model, and are used in calculating the MSE.

df used to describe the no of independent ways in which a system can move on change.

What is t-test for testing hypothesis for linear regression

In regression analysis, the t-test is used to determine whether the regression coefficients for each predictor variables are significantly different from zero. A coefficient that is significantly different from zero indicates that there is a relationship between the predictor and response variables, and that the predictor can be used to predict the response.

The t-test is calculated as:

$$t = \frac{b_i}{SE(b_i)}$$ where b_i is the coefficient for the i-th predictor variable and $SE(b_i)$ is the standard error of the coefficient.

The t-statistic is then compared to a critical value from the t-distribution, with a specified level of significance, to determine whether the coefficient is significantly different from zero. For example, if the level of significance is 0.05 and the t-statistic for a predictor variable is greater than 2 (or less than 2), then the coefficient for that variable is considered to be significantly different from zero.

In statistic, a critical value is the value of a test statistic that is used to determine the significance of the observed results. In a hypothesis test, the critical value is compared to the calculated value of the test statistic to determine whether the null hypothesis should be rejected in favour of the alternative hypothesis.

For example, in a t-test for testing the significance of regression coefficients in a linear regression model, the critical value is the value of the t-statistic that is used to determine whether a coefficient is significantly different from zero. The critical value is determined by the level of significance chosen for the hypothesis test (e.g. 0.05), the degrees of freedom for the test, and the type of alternative hypothesis (one-sided or two-sided).

In summary, a critical value is the value of a test statistic that is used to determine the significanee of the observed results in a hypothesis test. It is compared to the calculated value of the test statistic to determine whether the null hypothesis should be rejected in favor of the alternative hypothesis.

In summary, the t-test is used to test the hypothesis that the regression coefficients for each predictor variable are significantly different from zero, and to determine whether the predictor variables can be used to accurately predict the response variable.

Here is an example of a t-test for testing the significance of regression coefficients in a linear regression model with two predictor variables (x_1, x_2) and one response variable (y):

x_1	x_2	x_3
1	2	5
2	3	7
3	4	9
4	5	11

The regression equation for this model is: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ where b_0, b_1 and b_2 are the regression coefficients. Assume that the estimated values for the coefficients are $b_0 = 1$, $b_1 = 2$ and $b_2 = 3$.

The standard errors for the coefficients can be calculated using the following formulas:

$$SE(b_0) = \text{sqrt}\left(SSE / (n-p-1) \times \left(1/n + \mu^2 / SS_{xx}\right)\right)$$
$$SE(b_1) = \text{sqrt}\left(SSE / (n-p-1) \times \left(1/SS_{xx}\right)\right)$$
$$SE(b_2) = \text{sqrt}\left(SSE / (n-p-1) \times \left(1/SS_{xx}\right)\right)$$

where SSE is the sum of squares error, n is the total number of observations, p is the number of predictor variables, μ is the mean of the observed response values and SS_{xx} is the sum of squares of the predictor variables.

For example, the standard errors for the coefficients are:

$$SS(b_0) = \text{sqrt}\left(10/2 \times \left(1/4 + (8.5)^2/40\right)\right) = 0.45$$
$$SS(b_1) = \text{sqrt}\left(10/2 \times \left(1/40\right)\right) = 0.15$$
$$SS(b_2) = \text{sqrt}\left(10/2 \times \left(1/40\right)\right) = 0.15$$

The t-statistics for the coefficients can then be calculated as:

$$t(b_0) = 1 / 0.45 = 2.22$$
$$t(b_1) = 2 / 0.15 = 13.33$$
$$t(b_2) = 3 / 0.15 = 20.00$$

The critical value for the t-statistic, with a level of significance of 0.05 and two degrees of freedom ($df = n-p-1$) is 2.92 for a one-sided alternative hypothesis and 2.01 for a two-sided alternative hypothesis.

Since the calculated t-statistics for b_1 and b_2 are greater than the critical value, the coefficients for these predictor variables are considered to be significantly different from zero. This indicates that x_1 and x_2 can be used to predict the response variable y.

In summary, the t-test is used to determine whether the regression coefficients for each predictor variable are significantly different from zero and to determine whether the predictor variables can be used to accurately predict the response variable. The t-statistics for the coefficients are calculated and compared to the critical value for the t-distribution to determine their significance.

What is one sided alternative hypothesis and two sided alternative hypothesis ?

In a hypothesis test, the alternative hypothesis is a statement that is tested against the null hypothesis. The alternative hypothesis is usually the opposite of the null hypothesis, and is typically used to express the research hypothesis or the hypothesis of interest.

The alternative hypothesis can be one-sided or two-sided. A one-sided alternative hypothesis is a statement that specifies the direction of the difference or relationship between the variables, while a two-sided alternative hypothesis is a statement that does not specify the direction of the differences or relationship.

For example, in a t-test for testing the significance of regression coefficients in a linear regression model, a one-sided alternative hypothesis might be "the coefficient for variable x_1 is significantly greater than zero", while a two-sided alternative hypothesis might be "the coefficient for variable x_1 is significantly different from zero."

In summary, the alternative hypothesis is a statement that is tested against the null hypothesis in a hypothesis test. The alternative hypothesis can be one-sided or two-sided, depending on whether it specifies the direction of the difference or relationship between the variables.

What is F-statistic for estimating significance in regression?

In regression analysis, the F-statistic can be used to evaluate the overall significance of the regression model. This is done by conducting an F-test, which is a hypothesis test that compares the mean square of the regression (MSR) to the mean square error (MSE).

The null hypothesis for the F-test is that all of the predictor variables in the model are not significantly related to the response variable. If the p-value for the F-statistic is less than a pre-determined significance level (e.g., 0.05), then the null hypothesis can be rejected and it can be concluded that at least one of the predictor variables is significantly related to the response variable.

In other words, the F-statistic is used to determine whether the regression model as a whole is significantly related to the response variable. This is useful because it allows you to evaluate the overall significance of the model, rather than just the significance of the model, rather than just the significance of individual predictor variables. By using the F-statistic, you can determine whether the model is a good fit for the data and whether it can be used to make reliable predictions about the response variable.

What is the p-value ?

The p-value is a probability measure that is used in statistical hypothesis testing to determine the significance of the observed results. It is the probability of obtaining a sample statistic that is at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

For example, suppose a researcher is conducting a study to examine the effectiveness of a new drug in treating a certain condition. The researcher collects data on the effectiveness of the drug in a sample of individuals and calculates a sample statistic (e.g. the mean improvement in symptoms). The researcher then compares the sample statistic to the hypothesized value under the null hypothesis (e.g. no improvement in symptoms) and calculates the p-value.

If the p-value is less than a pre-determined significance level (e.g., 0.05), then the null hypothesis can be rejected and it can be concluded that the observed results are statistically significant. This means that the sample statistic is unlikely to have occured by chance and that the drug is likely to be effective in treating the condition. On the otherhand, if the p-value is greater than the significance level, then the null hypothesis cannot be rejected and it cannot be concluded that the observed results are statistically significant.

In summary, the p-value is a measure of the statistical significance of the observed results in a hypothesis test. It is the probability of obtaining a sample statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. If the p-value is less than the pre-determined significance level, then the null hypothesis can be rejected and it can be concluded that the observed results are statistically significant.

What is null hypothesis and alternative hypothesis in f-statistic?

In regression analysis, the null hypothesis for the F-statistic is that all of the predictor variables in the model are significantly related to the response variable. The alternate hypothesis is that at least one of the predictor variables is significantly related to the response variable.

The F-statistic is calculated by dividing the mean square of the regression (MSR) by the the mean square error (MSE). If the F-statistic is large, it indicates that the model explains a significant amount of the variation in the response variable, and the null hypothesis can be rejected in favor of the alternate hypothesis. This means that at least one of the predictor variables is significantly related to the response variable.

To summarize, the null hypothesis for the F-statistic in regression analysis is that all of the predictor variables are not significantly related to the response variable, while the alternate hypothesis is that at least one of the predictor variables is significantly related to the response variable. The F-statistic is used to evaluate the overall significance of the regression model by compairing the mean square of the regression to the mean square error. If the F-statistic is large, it indicates that the model is significantly related to the response variable and the null hypothesis can be rejected in favor of the alternative hypothesis.

Assumptions of linear regression

Linearity assumption

The linearity assumption in linear regression refers to the assumption in linear regression refers to the assumption that the relationship between the dependent variable and the independent variables is linear

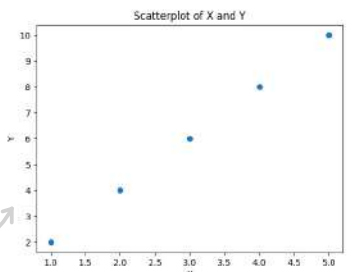
This means that if you plot the data on a graph, the points should form a straight line. If the points do not form a straight line, then the relationship between the variables is non-linear, and a different type of statistical model might be more appropiate.

For example, suppose you want to model the relationship between a person's age (the independent variable) and their income (the dependent variable). If you plot the data on a graph, and the points form a straight line, then you can use linear regression to model the relationship age and income. However, if the points do not form a straight line, then the relationship between age and income is non-linear, and you might need to use a different type of statistical model.

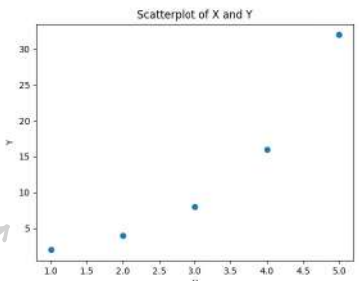
In summary, the linearity assumption in linear regression refers to the assumption that the relationship between the dependent variables and the independent variables is linear. If this assumption is not met, then a different type of statistical model might be more appropiate.

Suppose we have the following data:

X	Y
1	2
2	4
3	6
4	8
5	10



X	Y
1	2
2	4
3	8
4	16
5	32



Assumptions Test

visual inspection

One approach is to create a scatterplot of the predictor variables and the response variable, and visually inspect the plots to see if there is a linear relationship. If the relationship appears to be non-linear, you can try transforming predictor variables and re-plotting the scatterplots to see if the relationship becomes more linear.

Correlation matix

Another approach is to use statistical tests to assess the linearity assumption. One common method is to use the Pearson correlation coefficient to measure the strength of the linear relationship between the predictor variables and the response variable. A correlation coefficient of 0 indicates no linear relationship, while a coefficient of 1 indicates a perfect linear relationship.

Linearity test

You can also use a linearity test, such as the Goldfeld-Quandt test, to formally test the linearity assumption.

Remedies

Trasformation of vars.

Transforming the predictor variables: If the relationship between the predictor variables and the response variable appears to be non-linear, you can try transforming the predictor variables to see if it improves the linearity assumption.

For example, you can try taking the log, square root, or square of the predictor variables.

Interaction terms

Adding higher-order terms or interaction terms: You can try adding higher-order terms or interaction terms to your model to see if it improves the linearity assumption.

For example, if you have a predictor variable that appears to have a non-linear relationship with the response variable, you can try adding a quadratic term (x^2) to your model to see if it improves the fit.

Using a different

Using a different model: If the linearity assumption cannot be satisfied using the above methods, you may need to consider using a different type of model, such as a nonlinear regression model or a generalized linear model.

It's important to note that it's not always necessary to satisfy the linearity assumption in order to build a good predictive model. However, if the linearity assumption is significantly violated, it can lead to biased and inaccurate estimates of the model parameters, so it's worth considering these remedies to see if they can improve the model fit.

For example, if the linearity assumption is violated and the model includes a predictor variable that has a non-linear relationship with the response variable, the estimated coefficients for that predictor variable may not accurately reflect the true relationship between the variables. This can lead to incorrect predictions and a poorer fit of the model to the data.

Independence assumption

The independence assumption for linear regression refers to the assumption that the errors (the differences between the predicted values and the observed values) are independent of each other. In other words, it assumes that the errors are not related to one another and are not influenced by previous errors.

Violating the independence assumption can affect the validity of the model and the conclusions drawn from it. If the errors are not independent, it can result in biased estimates of the model coefficients and inflated standard errors, which can affect the precision and accuracy of the model.

One way to test for independence is to plot the residuals (the differences between the predicted values and the observed values) against the predicted values. If the residuals are randomly scattered around a horizontal line, it suggests that the errors are independent of the predicted values. This can provide some evidence that the independence assumption has not been violated.

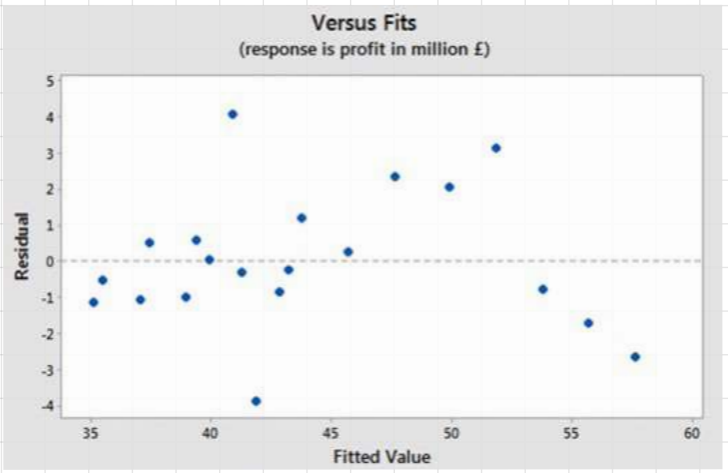
Another way to test for independence is to use a statistical test, such as the Durbin-Watson test. This test compares the autocorrelation of the residuals to a known value. If the test statistic is close to the known value, it suggests that the errors are independent.

It's important to note that these tests are not foolproof and it is still possible for the independence assumption to be violated even if the tests do not detect any issues. It is always a good idea to carefully examine the data and consider the context in which the data was collected to ensure that the independence assumption has not been violated.

An example of the independence assumption being violated in a linear regression model might be a study examining the relationship between the amount of time a student spends studying and their exam performance. If the researchers collected data on the amount of time each student spent studying and their corresponding exam score, but the students were allowed to collaborate and share notes, this could violate the independence assumption. In this case, the errors (differences between the predicted exam scores and the actual exam scores) would not be independent, because the students' exam scores would be influenced by each other's study habits and notes.

Another example might be a study examining the relationship between the price of a product and the number of units sold. If the data was collected over a period of time, and the price of the product was changed during that period, the errors (differences between the predicted number of unit sold and the actual number of units sold) would not be independent. This is because the number of units sold would be influenced by the price of the product, which changed over time.

In both of these examples, violating the independence assumption could lead to incorrect conclusions about the relationship between the predictor and outcome variables, and could result in inaccurate predictions or estimates. It is important to carefully consider the context in which the data was collected and to check for violations of the independence assumption before drawing conclusions from the data.



Homoscedasticity

Homoscedasticity is a statistical assumption that states that the variance of the error term (also known as residuals) in a regression model is constant across all values of y predictor variables. In other words, it assumes that the amount of error in the model is the same regardless of the value of the predicton variables. This is an important assumption because it allows us to use statistical techniques that rely on the assumption of constant variance, such as t-tests and F-tests, to test hypotheses about the relationships between the predictor and response variables in the model.

In layman's terms, homoscedasticity can be thought of as the assumption that the spread or dispersion of the residuals (errors) is constant across all values of the predictor variables. If the varience of the residuals is not constant, it is said to be heteroscedastic, which can lead to problems when using certain statistical techniques to analyze the model.

It's important to note that homoscedasticity is not always present in real-world data, and it is often necessary to check for and correct for heteroscedasiticity in order to properly analyze a regression model. There are vamous techniques lhal can be used to check for and adress heteroscedasiticity, such as transforming the data or using weighted least squares regression.

Normality Assumption

The normality assumption for linear regression refers to the assumption that the errors (the differences between the predicted values and observed values) are normally distributed. In other words, it assumes the errors follow a bell shaped curve, with most the errors clustered around the mean and fewer errors at the extreme ends of the distribution.

Violating the normality assumption can affect the validity of the model and the conclusions drawn from it. If the errors are not normally distributed, it can result in biased estimates of the model coefficients and inflated standard errors, which can affect the precision and accuracy of the model.

One way to test for normality is to plot the residuals (the differences between the predicted values and the observed values) and look for a bell-shaped curve. If the residuals are normally distributed, it suggests that the normality assumption has not been violated.

Another way to test for normality is to use a statistical test, such as the Anderson-Darling test or the Shapiro-Walk test. These tests compare the distribution of the residuals to a normal distribution and provide a p-value, which can be used to determine whether the normality assumption has been violated.

No Multicollinearity Assumption

In the context of multiple linear regression, the no multicollinearity assumption refers to the assumption that there is no strong linear relationship between the predictor variables. In other words, it assumes that the predictor variables are not highly correlated with each other. This is an important assumption because it allows us to interpret the coefficients of the predictor variables in the regression model as measures of the unique effect of each predictor on the response variable, holding all other predictors constant.

In layman's terms, the no multicollinearity assumption can be thought of as the assumption that the predictor variables are independent of each other, or that they do not have a strong linear relationship. If the predictor variables are highly correlated, it can be difficult to disentangle their individual effects on the response variable, which can make it difficult to interpret the results of the regression analysis.

It's important to note that multicollinearity is a common problem in regression analysis, and it is often necessary to check for and address multicollinearity in order to properly analyze a multiple linear regression model. There are various techniques that can be used to check for and address multicollinearity, such as examining the variance inflation factor (VIF) or using ride regression or LASSO regression.

No Multicollinearity Assumption Example

Suppose you are a marketing research analyst and you want to understand the factors that influence sales of a particular product. You collect data on several potential predictor variables, including the price of the product, the advertising budget for the product, and the number of stores that carry the product. You then fit a multiple linear regression model to predict sales as a function of these predictor variables.

In this, the no multicollinearity assumption would be important because you want to be able to interpret the coefficients of the predictor variables as measures of their unique effect on sales, holding all other predictors constant. If there is a strong linear relationship between the predictor variables (e.g., if the advertising budget is highly correlated with the number of stores that carry the product), it can be difficult to disentangle their individual effects on sales, which can make it difficult to interpret the results of the regression analysis.

To check for multicollinearity in this case, you might examine the correlation matrix of the predictor variables and look for strong correlations between the variables. You could also compute the variance inflation factor (VIF) for each predictor variable, which measures the extent to which multicollinearity is present in the model. If the VIF values are all below a certain threshold (e.g. 5 or 10), it is likely that multicollinearity is not a significant problem in the model.