## use case Regression - Numeric prediction

user case = hdb flat price prediction

independent features = noof rooms, sft, reno, nearmrt

dependent feature( predicted variable) = price

1. Simple Linear Regression

only one independent feature = sft = one X feature -

only one dependent FEATURE = price = One Y feature -

2. Multiple Linear regression

many independent features = sft ,reno, near\_mrt, no\_of\_rooms = many X features

only one dependent FEATURE = price = One Y feature

. Simple Linear Regression

Goal = take the independent feature value and predict the dependent feature value

Relationship between independent feature value and dependent feature value

Correlation : strength of two variables

Regression : how independent decides/effects the dependent , model fitting

Relationship between independent feature value and dependent feature value ( statistical relationship ) ( blk 616 senja road)



|  |  |
| --- | --- |
| Area in SFT( Square feet) independent feature value | Flat Price dependent feature value |
| 1500( x1) | 150000(y1) |
| 2500(x2) | 300000(y2) |
| 1800 | ? y |



Y = mx +c

M = slope

C = intercept

X = independent feature value

Y = dependent feature value

M = slope of straight line

M = y2-y1/x2-x1

= 300000 – 150000 / 2500-1500

= 150000/ 1000

= 150

Y = mx + c

Y = predicted flat price

X = area in sft ( 1800)

M = 150

If Y = mx + c

C = Y – mx, use if for all data points

|  |  |
| --- | --- |
| 1500( x1) | 150000(y1) |
| C = 150000 – 150\* 1500  C = -75000 |  |

C = Y – mx, use if for all data points

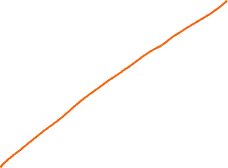
|  |  |
| --- | --- |
| 2500(x2) | 300000(y2) |
| C = 300000 – 150\*2500  = -75000 |  |

Y = mx + c for my flat sft of 1800?

Y = 150\* 1800 +-75000

= 195000

Y= predicted price



Y precited = 195000

Actual = 220000

(Error)Difference between actual and predicted = 220000 – 195000 = $25000

Y precited = 195000

Actual = 165000

(Error)Difference between actual and predicted = 165000 – 195000 = -$30000

To calculate the error of all the rows ( square of all errors)

* Least square method

Y = mx + c

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y( acutal Y) | Predicted y ( yhat)y^ | y-y^ | (y-y^)2 |
| 3 | 1 | Y^ = 3m +c | 1-3m+c | (1-3m+c)2 |
| 5 | 5 | Y^ = 5m+c | 5-5m+c | (5-5m+c) 2 |
| 7 | 8 | Y^=7m+c | 8-7m+c | (8-7m+c) 2 |
|  |  |  |  | Sum of all squares  (1-3m+c)2 + (5-5m+c) 2 + (8-7m+c) 2 |

1. 3m + c)2

(A+b)2 = a2 + 2ab +b2

(A+b+c)2  = a2 + b2 + c2 + 2ab + 2bc + 2ac

A = 1

B = 3m

C = c

1. 3m + c)2  = 1 + 9m2 + c 2 – 6m +6mc + 2c

After all the recalculation we get m and we get c

We can go with y = mx + c

You may get the best fit line

If all this to put in programmatical ( traditional ) + sklearn

1. X features
2. Y feature
3. X mean
4. Y mean
5. X sum
6. Y sum
7. Sum(x\*y) n
8. X\*x
9. Square of
10. Slope . y2-y1/x2-x1
11. Intercept C = Y – mx
    1. C = ymean – m(slope) \* xmean

Once we get the slope and intercept , we go for the prediction of the Y for entire dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Actual Enginesize | Actual co2emission | Predicted y co2emission | Error = Actual co2emission- Predicted y co2emission | Error = (Actual co2emission- Predicted y co2emission)2 |
|  |  |  |  | Finally sum the error |

Multiple LR

Multiple Linear regression

Y = mx + c

B =c

M = 𝛽1

𝑦̂=𝑓(𝑥⃗)=𝛽⃗⋅𝑥⃗+𝑏=𝛽1𝑥1+𝛽2𝑥2+⋯+𝛽𝑛𝑥𝑛+𝑏

Many independent variables

many independent features = sft ,reno, near\_mrt, no\_of\_rooms = many X features

only one dependent FEATURE = price = One Y feature

mx = m1x1 + m2x2+ m3x3+m4x4+……mnxn. + b

x1 = independent feature 1/ independent column1

x2 independent feature 2/ independent column 2

x3 independent feature 3/ independent column 3

x4 independent feature 4/ independent column 4