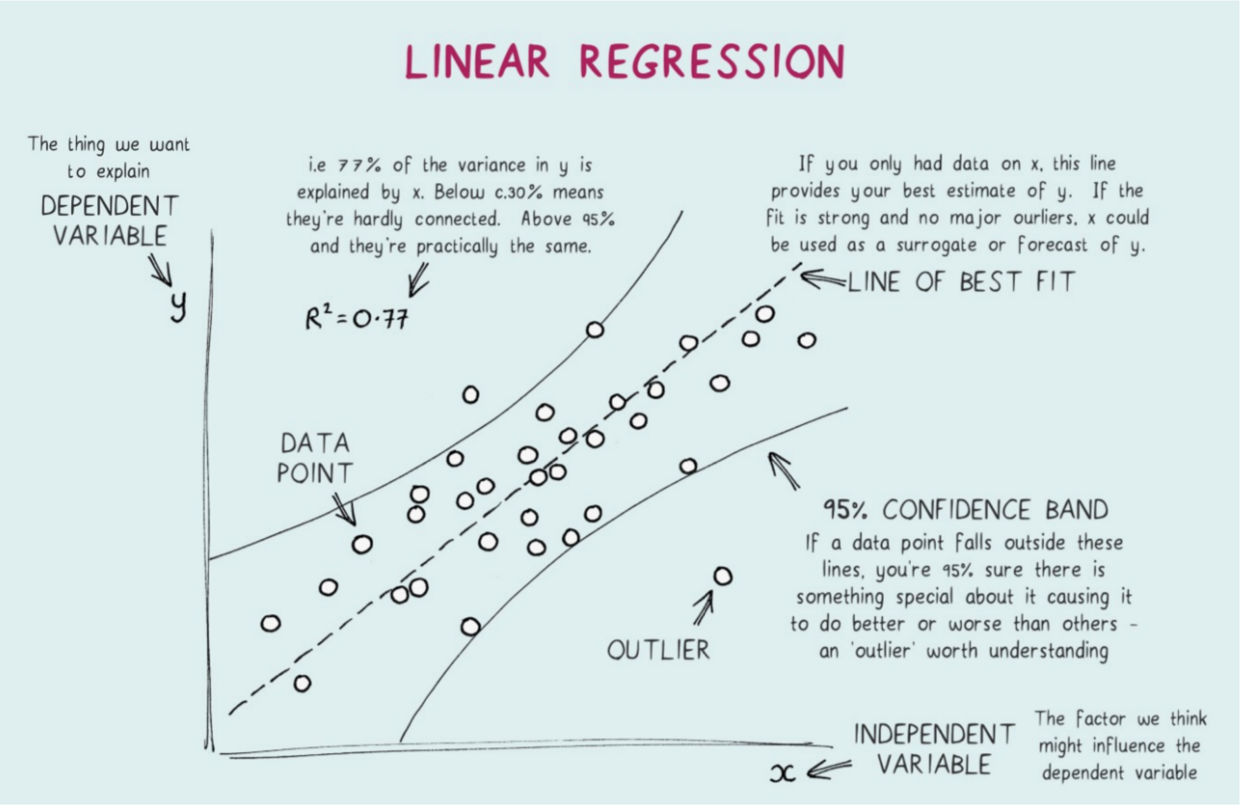
Linear Regression

What is Regression:

Regression is a statistical methodology where we could find the strength and relationship between X and Y, there can be a single predicting variable or Multiple predicting variables. The concept can be extended to Multiple variables as well.

Linear Regression AKA Least Squares:

Linear Regression model is trying to fit a linear line to the data. Next we will calculate the difference between the predicted points (which actually the linear line demonstrates) and the actual points. Our goal is to find a line which fits the data with the minimum sum of squares(errors). So the name 'LEAST SQUARES'


LINEAR REGRESSION
The thing we want to explain
DEPENDENT VARIABLE
y
i.e 77% of the variance in y is explained by x. Below c.30% means they're hardly connected. Above 95% and they're practically the same.
R^2=0·77
If you only had data on x, this line provides your best estimate of y. If the fit is strong and no major ourliers, x could be used as a surrogate or forecast of y.
LINE OF BEST FIT
DATA POINT
95% CONFIDENCE BAND
If a data point falls outside these lines, you're 95% sure there is something special about it causing it to do better or worse than others - an 'outlier' worth understanding
OUTLIER
INDEPENDENT VARIABLE
x
The factor we think might influence the dependent variable

Formulation:

Equation of a linear line would be : Y = mX + c

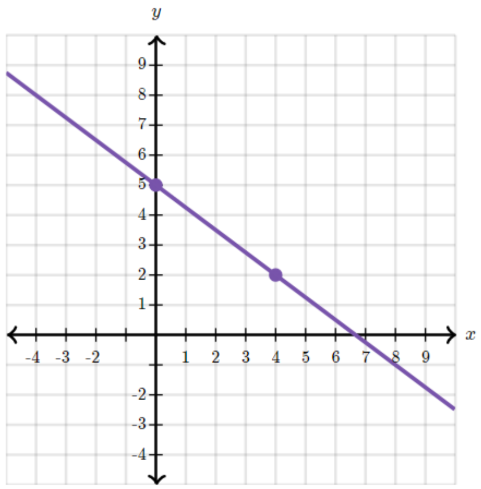
- Y - Dependent Variable
- X - Independent Variable
- m - Slope
- c - Intercept

What is Slope?

Slope can be told as a measure of steepness. It can also be measured as “rise(y-axis) over run(x-axis)”

m = dy/dx

The line appears to go through the points $(0, 5)$ and $(4, 2)$.



$$\text{Slope} = \frac{\Delta y}{\Delta x} = \frac{2-5}{4-0} = \frac{-3}{4}$$

What is Intercept?

Intercept is the point where the linear line cuts at the Y-axis. For example: In the above figure, the Linear Line is cutting the y-axis at y=5.
Hence Intercept=5.

Formulation for Linear Regression:

Consider Linear regression with two independent variables, the equation for the linear line would be

$Y = m_1X_1+m_2X_2+m_{0.}$

m_1 = Slope coefficient of X_1
m_2 = Slope coefficient of X_2

House_price = 2.42 * No_of_Bedrooms + 3.41 * No_of_amenities + 3.5

No_of_Bedrooms has slope coefficient of 2.42, which infers that the model changes by 2.42x for every one-unit change in No_of_Bedrooms

What will be the best regression line?

We can draw any line and calculate the metric. But the optimal line would be where the linear line also follows the relationship between the dependent and predicted variables.

It must try to cover as many data points as we drew over them and the distance of the data points should be minimum from the line.

Approach:

First we take the mean of training data and for every given y_i we find a difference between the mean and the actual predicted value.
Then we take the square of the difference values with mean
This will be noted as **Sum of Squared Errors (Total)**

Next, for every data point, with the best line we found, we calculate the y_i . We will take the difference between the predicted and actual values, And we will take squares of them.This will be called as **Sum of Squared Errors (Residuals)**

R^2 Error:

$$1 - \frac{SSE_{RES}}{SSE_{TOT}}$$

A measure which tells how well the line fits the data.

So Increased value of R-squared tells us, the model is great

RMSE:

Root Mean Squared Error - We gonna take Mean of sum of squared residuals and take square of it

Lower the RMSE, better the model

Assumptions to Linear regression model:

If the linear regression model doesn't work as expected. Check whether these assumptions for Linear models are met

1) Linear relationship between variables
2) Multicollinearity
3) Residual Independence (Time series data)
4) Homoscedasticity
5) Normality

The following article explains what to do when these assumptions are not met.
https://www.statology.org/linear-regression-assumptions/

Hands-on Practice