# MACHINE LEARNING

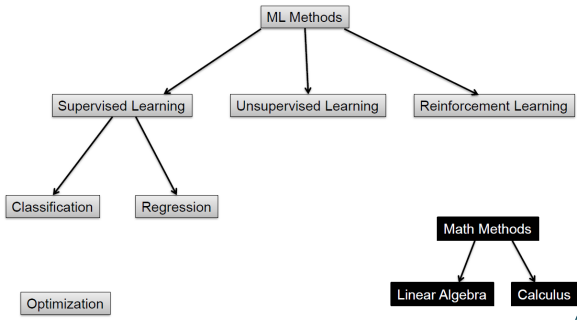
**Machine Learning** is the technique that figures out sorts of rules based on which the decisions were made on the historical data and uses the same rules to make decisions cleverly on the present data. It's not the robotic process of if-else, it's more of understanding all the measures involved to make the decision.

This type is called **supervised learning**. Where the decisions (**target**) were present in the historical data.

The target can be numeric **(regression)** or categorical **(classification)**
Example: numeric - amount given, categorical - loan given = yes or not.

**Unsupervised learning** - Looking after the data we set up that which set of data can be grouped together.

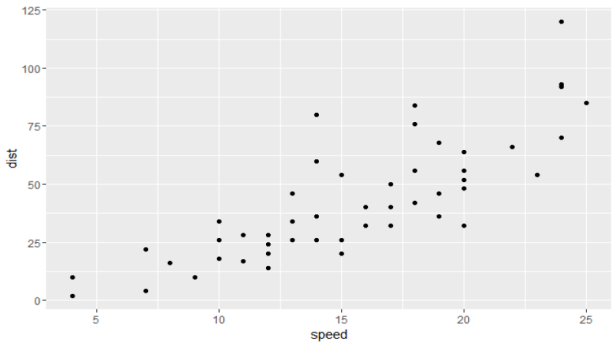
**Machine Learning Overview**




**Optimization**: This is the decision-making technique solved based on statistical analysis.

## **Simple Linear Regression**: USING LEAST SQUARE

- Independent variable (explanatory) – Speed (mph) – Plotted on X-axis
- Dependent variable (response) – Stopping distance(ft) – Plotted on Y-axis




**Our job is to create such a line that covers maximum points and has a minimum total error.**
**NOTE:** We are interested in minimizing the total error in the regression problem.

With respect to x, what is y present?
And with respect to x what we predict as y.
Now we want to reduce the error between actual and predicted

So, therefore, we want to minimise $y-\bar{y}$
Now the predicted can be both sides positive or negative so we take the **sum of square error.**

**Equation of a line:**
$y$**=mx+c,** where x is known and we need to find **slope m** and **intercept c** such that the total error should be minimum.

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

The value of *b*, the slope, that minimizes the SSE is given by

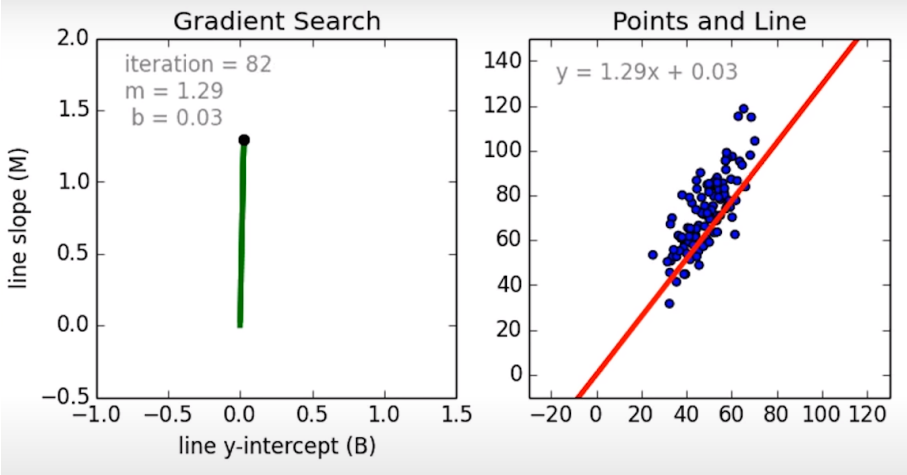
$$b = \frac{\sum((x_i - \bar{x})(y_i - \bar{y}))}{\sum(xi - \bar{x})^2}$$

(y = a + b$x$, m = b (c is used in us))

c=$\bar{y}$ - m$\bar{x}$

Once the slope and intercept is calculated, prediction is done.
example, x=1 m=3.2 and c=1.2, So y=3.2x+1.2
All the y are predicted and plotted with a line used to join the points.
This method of fitting the line is called the **least-squares** regression.
And we need such a line that has the least error.

SSE= ∑( yi - (mxi + c)^2 )




In the slope equation the numerator decides the sign of the equation, i.e tells how y moves as x moves. This is called **Covariance.**
Covariance measures the directional relationship. A positive covariance means that they move together while a negative covariance means they move inversely.

$$s_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$$

**But,** it doesn't tell about the strength or variation so the covariance is converted to a dimensionless quantity called **Correlation,** denoted by R.
The correlation coefficient, r, is the number between -1 to 1 tells us how well a regression line fits the data.

**Coefficient of determination: R^2**

It is the percentage of variation in the y variable that is explained by the x variable.
Ranges from 0 ( no linear relationship) to 1 (perfect linear relationship).

Distance actual - mean

vs

Distance predicted - mean

This is nothing but $R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$

**Hypothesis.**

Our hypothesis is for the population:
Null: x don't have an impact on y, as most of the time parameters don't have an impact on the result.
Alternate: x has an impact on y
So the slope becomes zero in 1st case. (population mean)
Therefore, Null - slope = 0
Alternate - slope != 0

Now, we calculate the slope, and we want to know how far it is from zero.
(How far Sample mean is from population mean)
So we check for z/t score, x-u/std error, where x is the slope ( sample mean calculated)
This tells us how many std deviation away from the mean
And p-value tells what is the value associated with the z-score
EXAMPLE:

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.717055011 |
| R Square | 0.514167888 |
| Adjusted R Square | 0.494734604 |
| Standard Error | 4.21319131 |
| Observations | 27 |

ANOVA

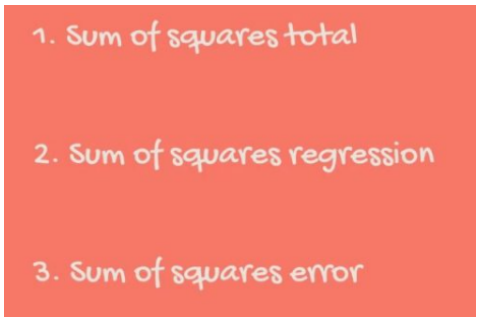
| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 469.6573265 | 469.6573265 | 26.4581054 | 2.57053E-05 |
| Residual | 25 | 443.7745253 | 17.75098101 | | |
| Total | 26 | 913.4318519 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 99.0%* | *Upper 99.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -4.154014573 | 2.447784673 | -1.697050651 | 0.102104456 | -9.195321476 | 0.88729233 | -10.97705723 | 2.669028089 |
| Big Mac Price ($) | 3.547427488 | 0.689658599 | 5.143744297 | 2.57053E-05 | 2.127049014 | 4.967805962 | 1.625048409 | 5.469806567 |

In the case of an intercept is greater than 0.05, we accept that the null hypothesis is true, c=0

Whereas for Big mac price we reject the null hypothesis, so slope(b) != 0, has a statistical impact on the result.

Rule of thumb: Having a smaller p-value means important, greater value not important.




*The **sum of squares total, denoted SST**, is the squared differences between the observed dependent variable and its mean.*
*The second term is the **sum of squares due to regression, or SSR**. It is the sum of the differences between the predicted value and the mean of the dependent variable*
*The last term is the **sum of squares error or SSE**. The error is the difference between the observed value and the predicted value.*

The Sum of squares in regression and the sum of squares of residuals/error is used as within and between groups to calculate the F score, which gives us F-significance.
If the f-significance is lower or up to 0.05 then the model has predictive power, if f-significance is more then the model doesn't have predictive power is better to go with average.
R squared == SSR/SST.

**Overall hypothesis for linear regression.**
Null: model doesn't have predictive ability, best estimator is average.
Alternate: The model has predictive ability better than average estimation.
**This is decided by looking at the F significance value.**

**Assumptions for linear regression.**

1. Model is linear
2. The error terms are independent
3. The error terms have constant variance ( homoscedasticity)
4. Residuals are normally distributed

Checking for normal distribution:

1. Plotting histogram
2. Plotting QQ plot - should show a straight line

Performing residual analysis

1. Fixing non-normality.
2. Fixing heteroscedasticity.

# MULTIPLE LINEAR REGRESSION

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.860872681 |
| R Square | 0.741101773 |
| Adjusted R Square | 0.715211951 |
| Standard Error | 11.96038667 |
| Observations | 23 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 8189.723012 | 4094.861506 | 28.62521631 | 1.35298E-06 |
| Residual | 20 | 2861.016988 | 143.0508494 | | |
| Total | 22 | 11050.74 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 57.35074586 | 10.00715186 | 5.73097587 | 1.31298E-05 | 36.47619286 | 78.22529885 |
| Area (sq ft) (x1) | 0.017718036 | 0.00314562 | 5.632605205 | 1.63535E-05 | 0.011156388 | 0.024279685 |
| Age of House (years) (x2) | -0.666347946 | 0.227996703 | -2.922620973 | 0.008417613 | -1.141940734 | -0.190755157 |

What is the equation?

$$\hat{y} = 57.35 + 0.0177Area - 0.666Age$$

Are the coefficients and the model significant?

Yes

All the concepts of simple linear regression work the same for the multiple as well.

# LINEAR REGRESSION USING GRADIENT DESCENT

## Mean Squared Error Function

1. Difference between actual value of y & predicted value of y

$$(y_i - \bar{y}_i)$$

actual value — predicted value

$$\bar{y}_i = mx_i + c$$

2. Square the Difference

$$(y_i - \bar{y}_i)^2$$

3. Find the mean of the squares

$$\frac{1}{n}\sum_{i=0}^{n}(y_i - \bar{y}_i)^2$$

$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$

## Gradient Descent Algorithm

Gradient Descent is an iterative optimization algorithm to find the minimum of a function. Here that function is our Loss Function.

### Step 1

Initially

m = 0

c = 0

L = Learning Rate = 0.0001

### Step 2

Calculate the partial derivative of loss function with respect to m & c

$$D_m = \frac{1}{n}\sum_{i=0}^{n} 2(y_i - (mx_i + c))(-x_i)$$

Partial Derivative wrt m, $D_m = \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - \bar{y}_i)$

Partial Derivative wrt c, $D_c = \frac{-2}{n}\sum_{i=0}^{n}(y_i - \bar{y}_i)$

## Step 3

Update current value of m & c using these equations:

$$m = m - L \times D_m$$

$$c = c - L \times D_c$$

## Step 4

Repeat Step 2 and Step 3 untill Loss = 0 (ideally)

L - Learning Rate

**PROS AND CONS:**

GRADIENT :

1. Works well with more number of features
2. Can be stuck in Local Minima
3. If not proper Learning Rate α, then it might not converge.

LEAST SQUARE:

1. No Learning Rate
2. Feature Scaling Not Necessary
3. Works really well when the Number of Features is less.
4. Is computationally expensive when the dataset is big.
5. Slow when the Number of Features is more.