# Introduction to ML & Linear Regression

# Agenda

- Introduction to Machine Learning
  - Supervised Learning
  - Unsupervised Learning
- Introduction to Supervised Learning
  - Regression
  - Classification
- Linear Regression using OLS
  - Simple Linear Regression
  - Multi Linear Regression

# Agenda

- Measure of Variation
  - SST - Sum of Square
  - SSR - Sum of Square of Regression
  - SSE- Sum of Square of Error
  - R- Square
  - Adjusted R-Square

# Introduction to Machine Learning

# Key Terminologies in the world of data

Artificial Intelligence Machine that mimics human intelligence & behaviour

Machine Learning Machine that learns from the data automatically w.r.t a particular task and performance measure

Deep Learning Subset of ML, learning happens via algorithms inspired by neural networks in human brain

Generative AI Subset of DL, creating new content from existing ones using advanced algorithms

# Key Terminologies in the world of data

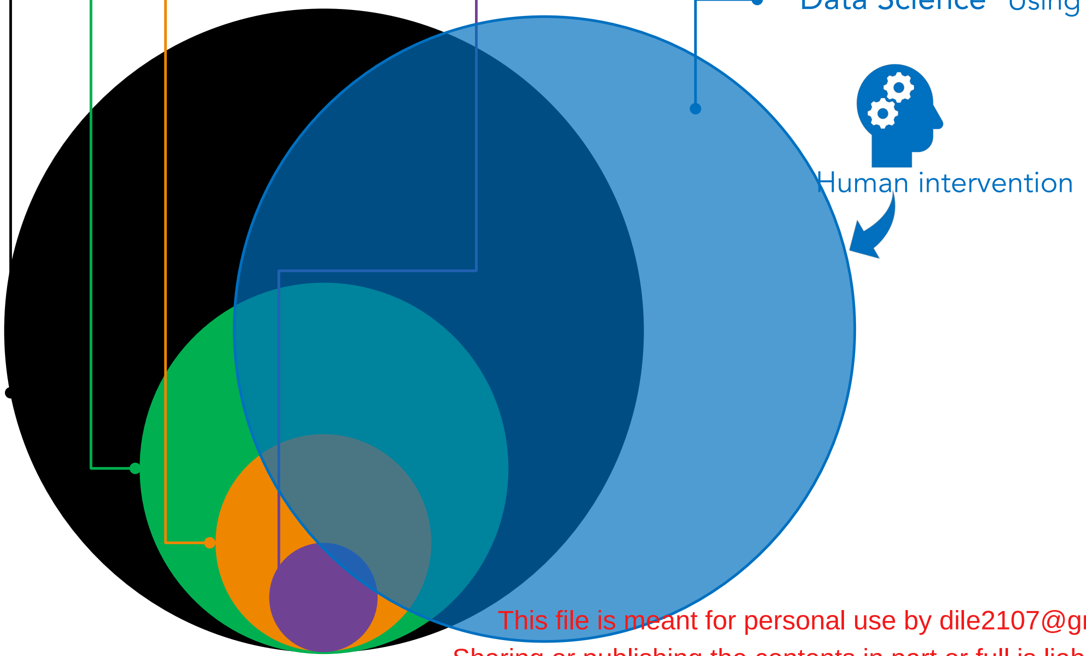
Artificial Intelligence Machine that mimics human intelligence & behaviour

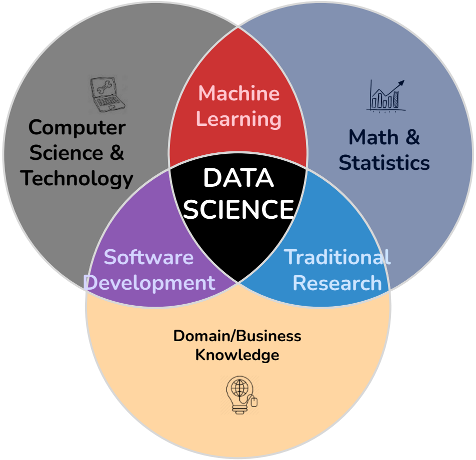
Machine Learning Machine that learns from the data automatically w.r.t a particular task and performance measure

Deep Learning Subset of ML, learning happens via algorithms inspired by neural networks in human brain

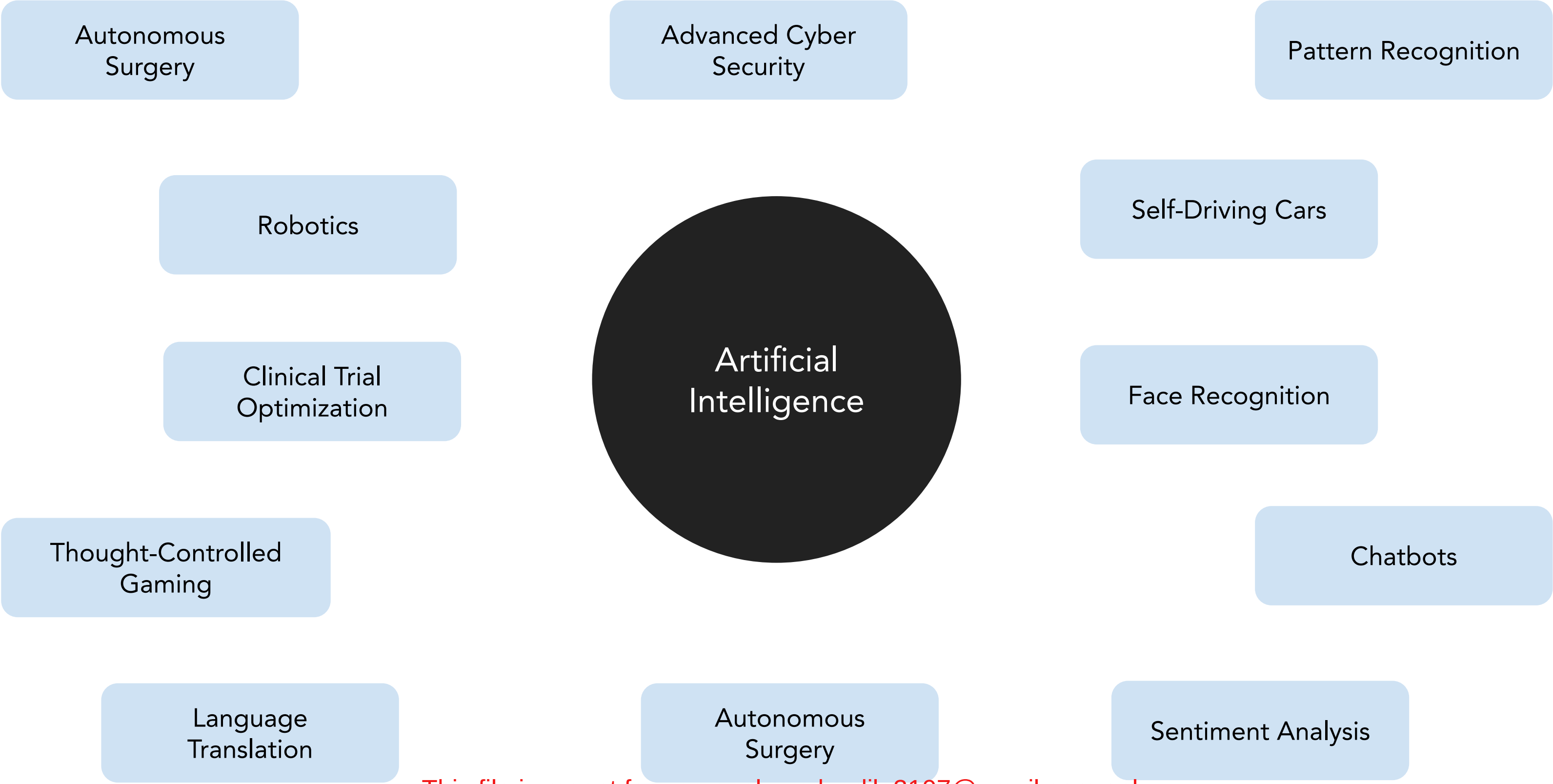
Generative AI Subset of DL, creating new content from existing ones using advanced algorithms

Data Science Using AI to enable data driven decision making



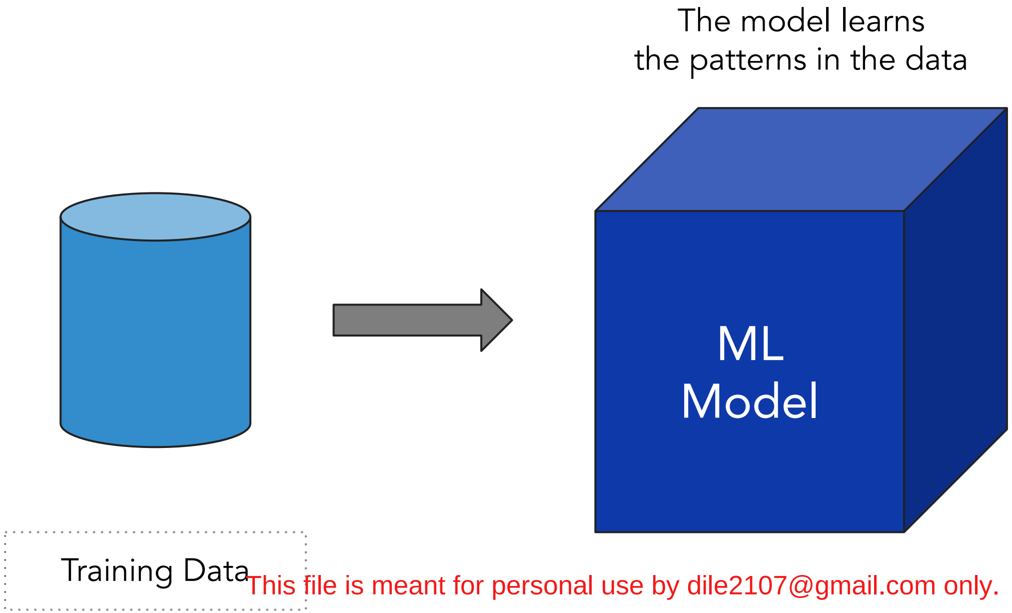



# Applications of AI

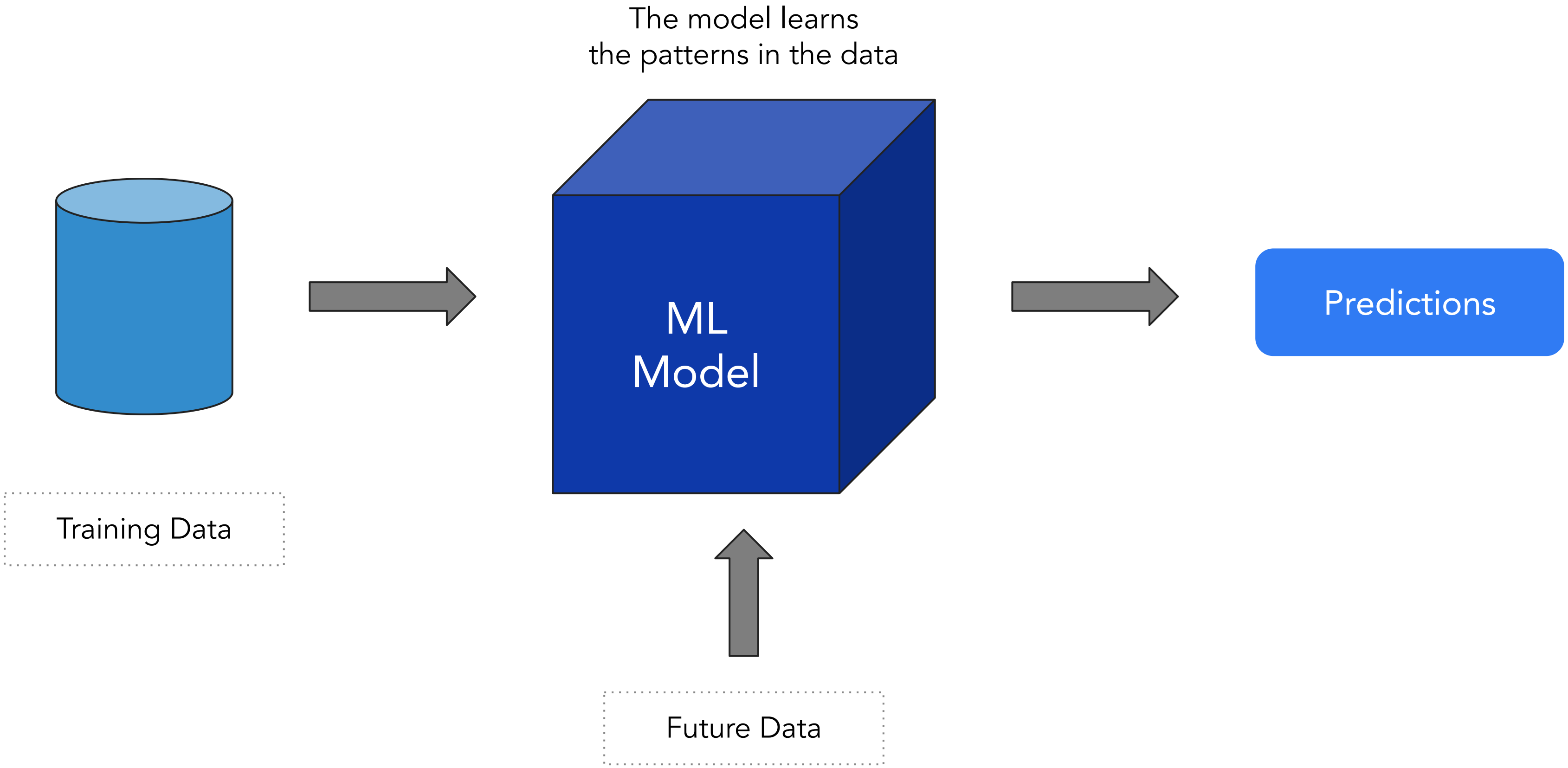


# Machine Learning

Machine Learning (ML) algorithms are developed to identify patterns within the dataset, develop an understanding, and predict output based on the understanding developed



# Machine Learning



# Types of Machine Learning

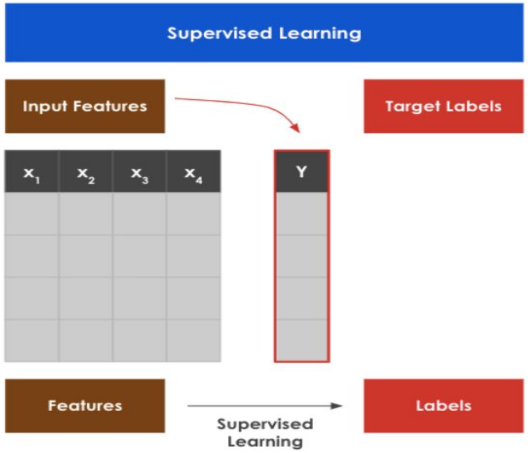
- Supervised Learning
  - These algorithms are trained on a labeled dataset, meaning they learn from input-output pairs. The supervised learning based algorithms try to learn the mapping from input variables to output variables, with the goal of being able to predict the correct output when given new input data.
- Unsupervised Learning
  - Unsupervised Learning algorithms are trained on unlabeled data, meaning they learn patterns and structures within the data without explicit guidance. In unsupervised learning, the algorithm explores the data to find hidden patterns or intrinsic structures, such as clusters or associations.
- Reinforcement Learning
  - Reinforcement learning is like teaching a robot to navigate through a maze by rewarding it with treats when it makes progress towards the goal and letting it try again when it makes mistakes.

In this course, you will learn Supervised and Unsupervised Learning

# Types of Machine Learning

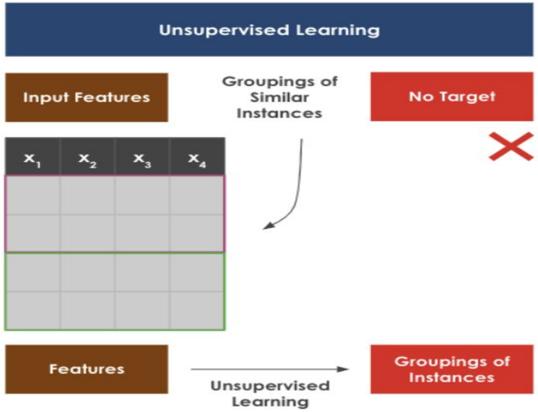
Supervised Learning

Supervised Learning algorithms need labeled data or target values for training.

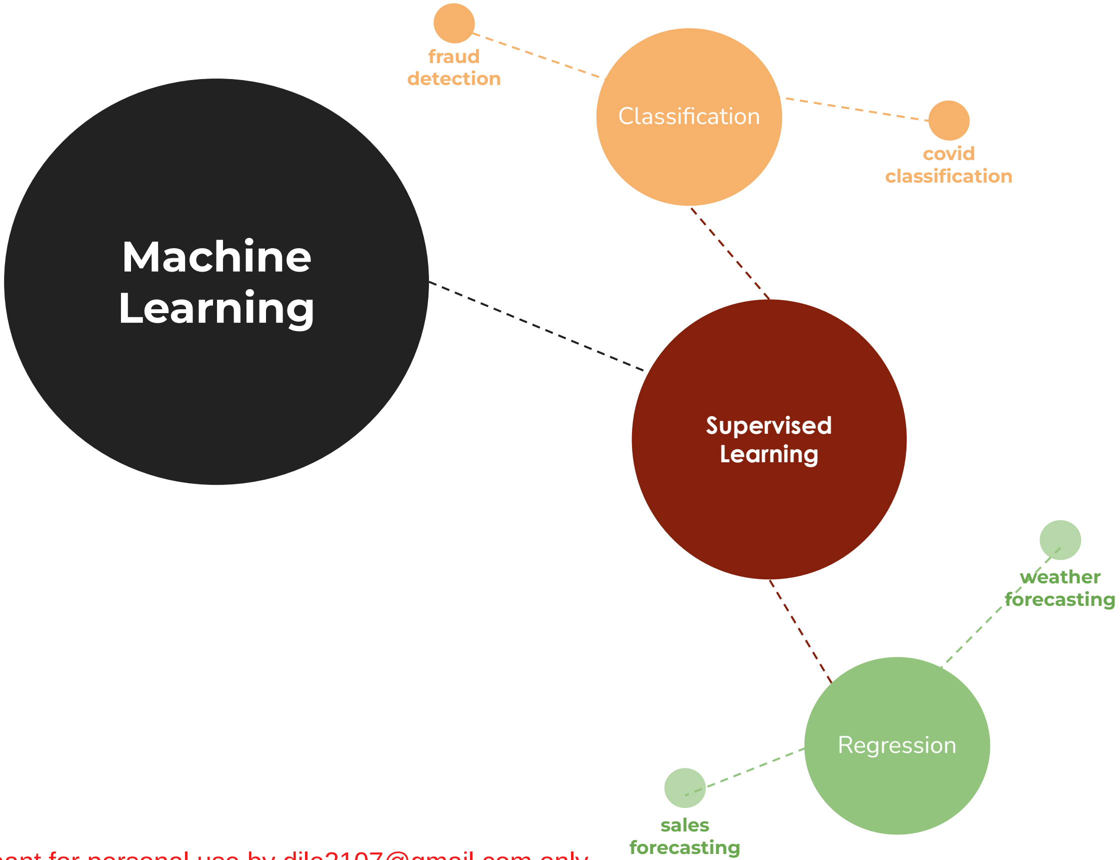



Unsupervised Learning

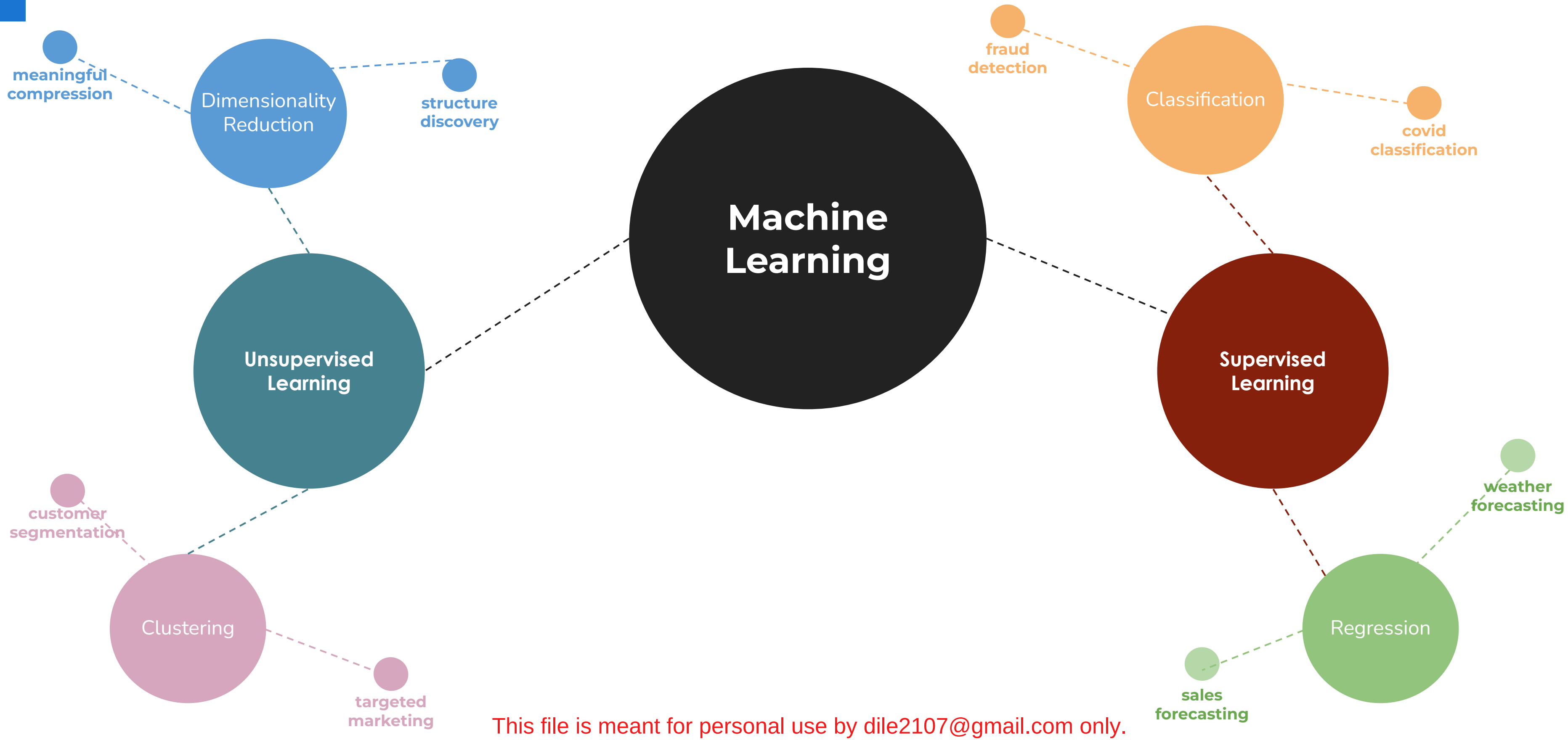
Unsupervised Learning algorithms do not need labeled data or target values.



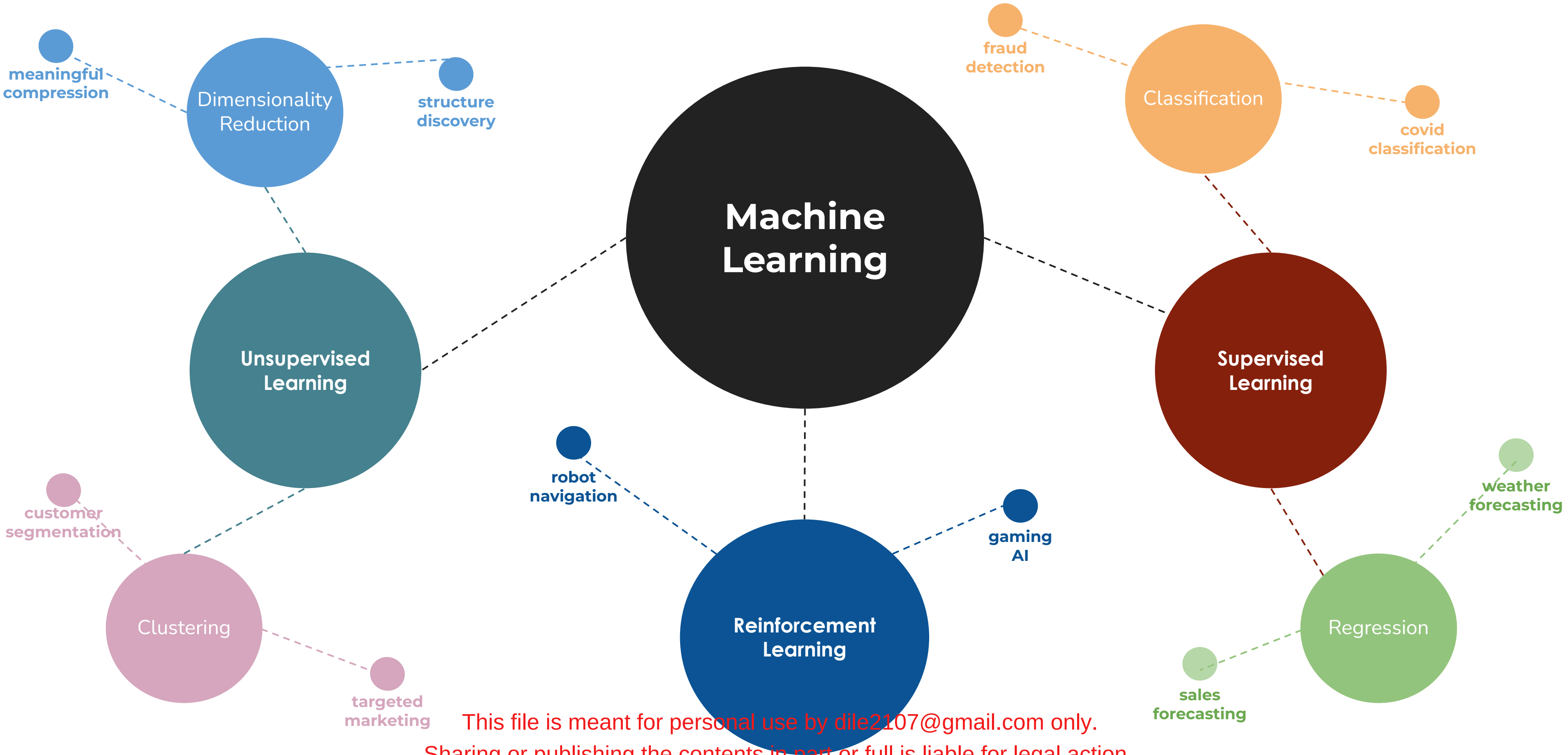
# Applications of ML - Supervised Learning



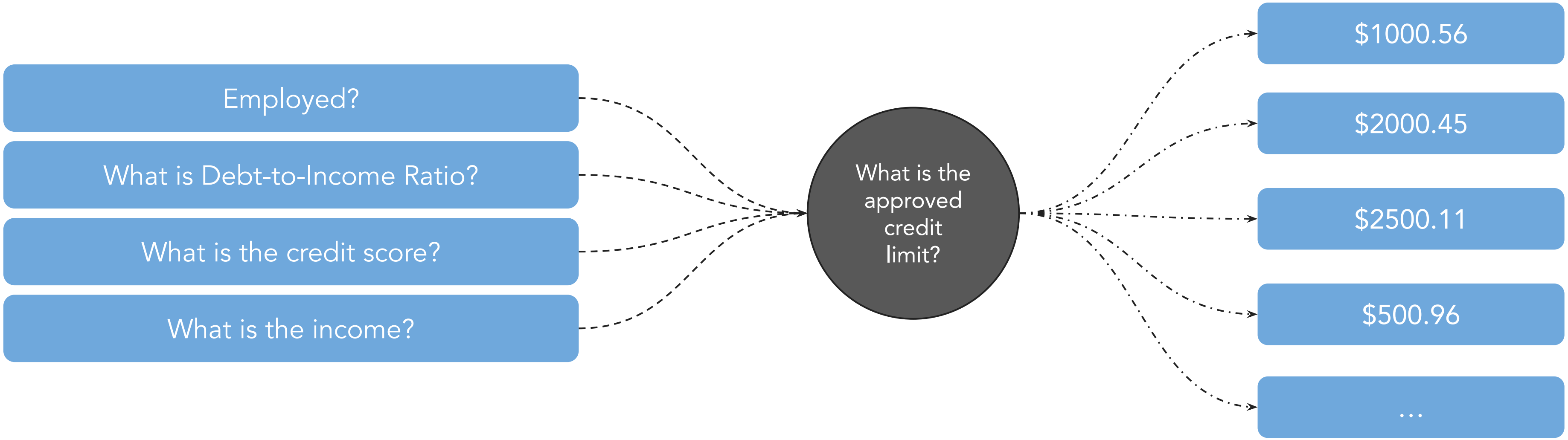
# Applications of ML – Unsupervised Learning



# Applications of ML - Reinforcement Learning



# Supervised Learning - Regression vs Classification

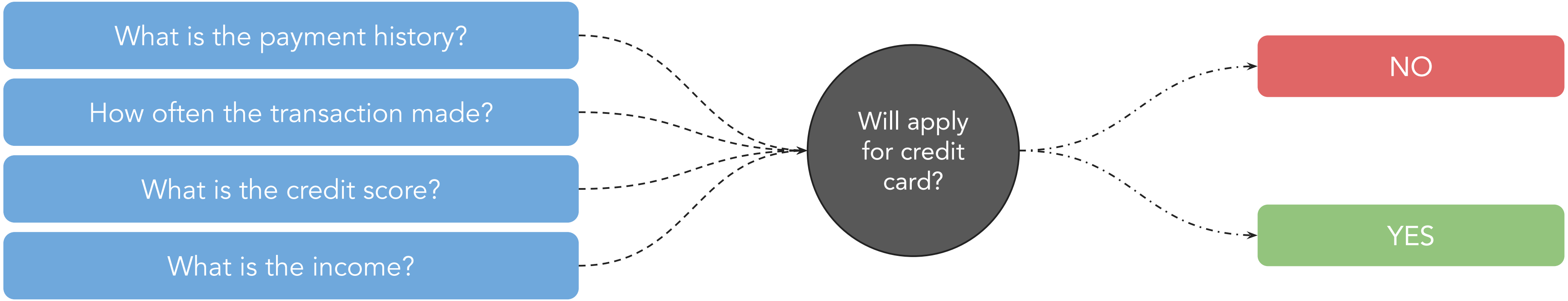



Continuous

Regression

The regressor learns the features of the input data and predict the limit

# Supervised Learning - Regression vs Classification

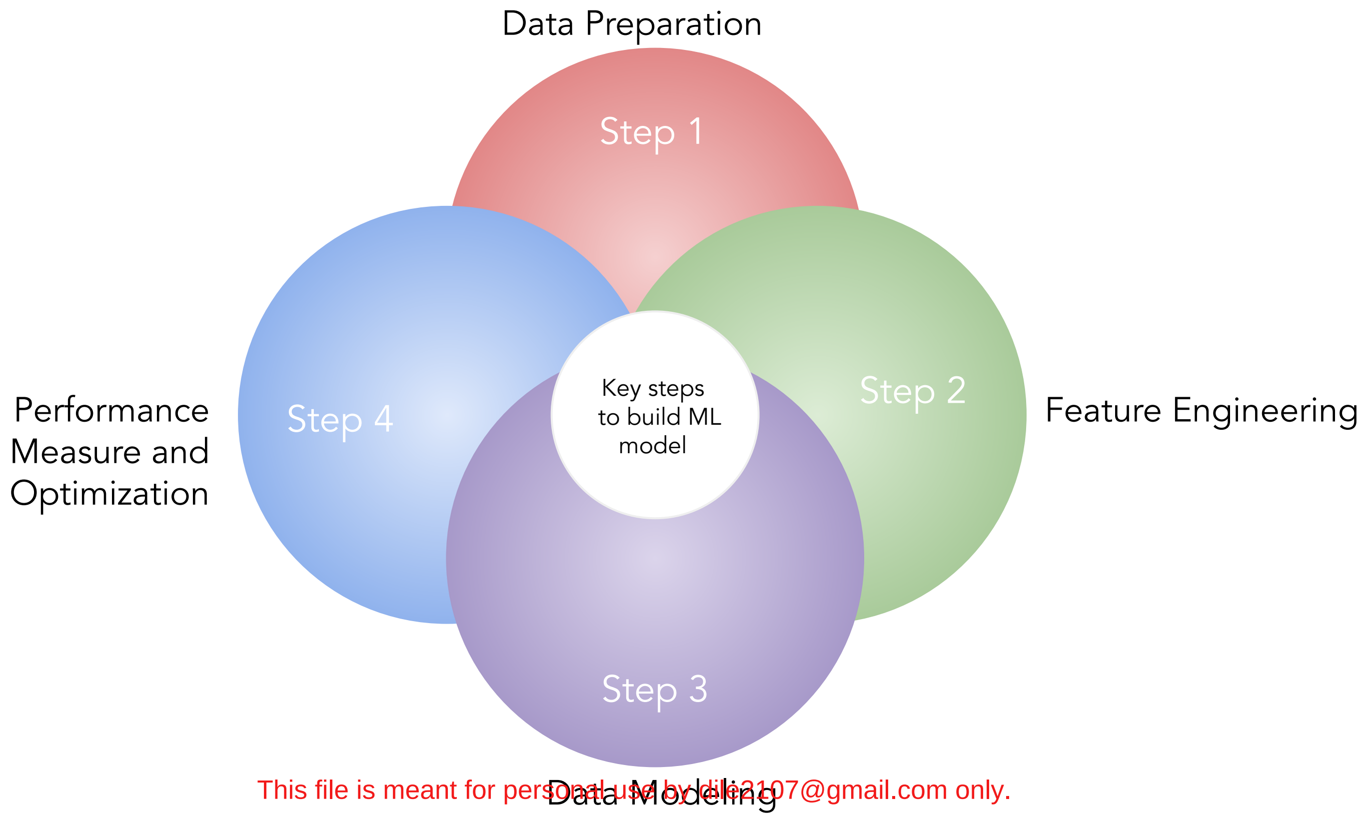



Category

Classification

The classifier learns the features of the input data to classify data into categories

# key steps to build ML model



# Interview Questions

- What is Machine Learning and what are its types?
- Difference between supervised and unsupervised models.
- Regression vs Classification.

# Simple Linear Regression

# Business problem: predict vehicle insurance premium

It is important for insurers to develop models that accurately forecast premium for car insurance.

These model estimates can be used to create premium tables that can assist to set the price of the premiums, depending on the expected treatment costs.

# Dependent variable

- The variable we wish to explain or predict
- Usually denoted by Y
- Dependent Variable = Response Variable = Target Variable
- Here ‘Insurance Premium’ is our target variable

# Independent variable

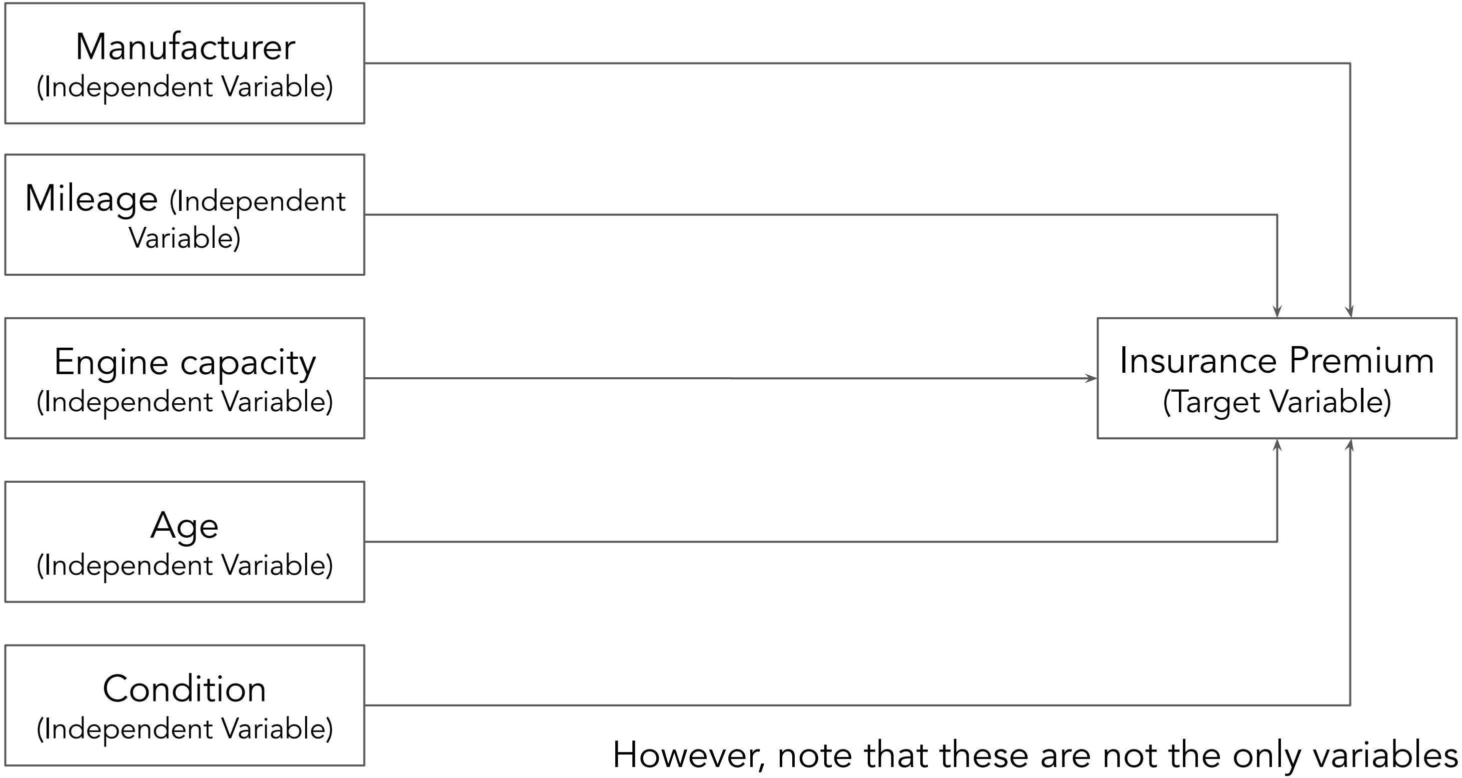
- The variables used to explain the dependent variable or used to help predicting the target variable

- Usually denoted by X

- Independent Variable = Predictor Variable = Features

- In our example, Age, Mileage and Condition of the car are the independent variables

# Variables that may contribute to insurance premium



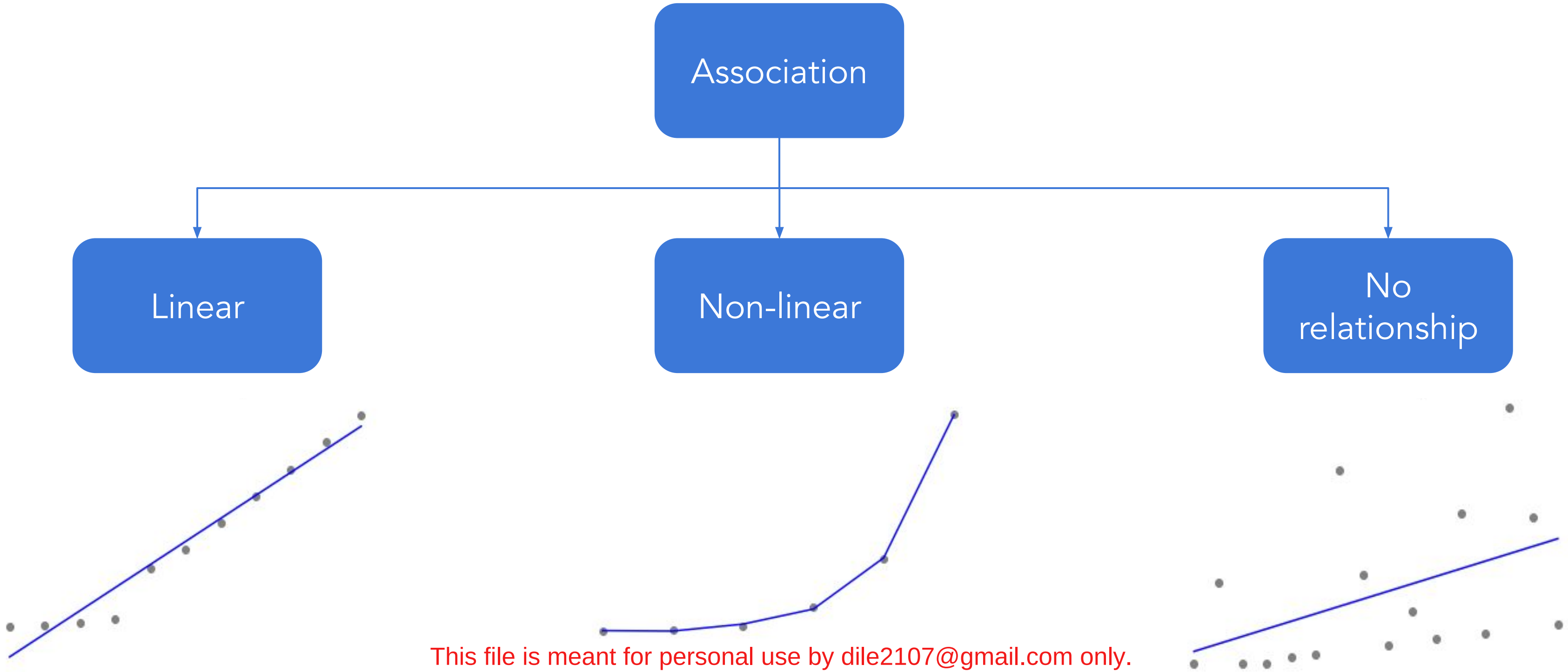

However, note that these are not the only variables considered. You may have some more in mind.

# Regression Analysis

# What is regression analysis?

- Regression analysis allows us to examine which independent variables have an impact on the dependent variable
- Regression analysis investigates and models the relationship between variables
- Determine which independent variables can be ignored, which ones are most important and how they influence each other
- We shall first see simple linear regression and then multiple linear regression

# Types of associations



# Simple linear regression

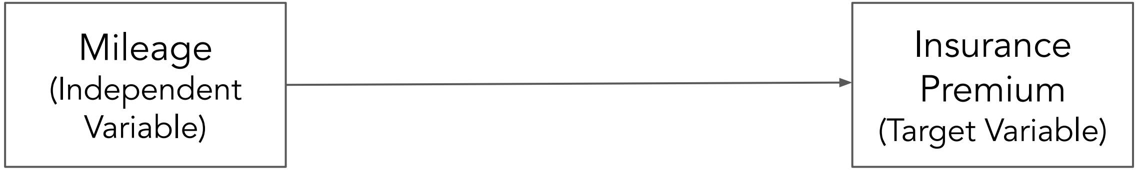
A simple linear regression model (also called bivariate regression) has one independent variable X that has a linear relationship with the dependent variable Y

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x + \boldsymbol{\varepsilon}$$

$\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are the parameters of the linear regression model.

# Variable that contributes to insurance premium

Let us consider impact of a single variable for now.




We say, that only mileage decides what the insurance premium should be.

# Data

Let us consider the following data.

| Mileage | Premium (in dollars) |
|---|---|
| 15 | 392.5 |
| 14 | 46.2 |
| 17 | 15.7 |
| 7 | 422.2 |
| 10 | 119.4 |
| 7 | 170.9 |
| 20 | 56.9 |
| 21 | 77.5 |
| 18 | 214 |
| 11 | 65.3 |
| 7.9 | 250 |
| 8.6 | 220 |
| 12.3 | 217.5 |
| 17.1 | 140.88 |
| 19.4 | 97.25 |

# Linear regression line

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x + \boldsymbol{\varepsilon}$$

$y$ = set of values taken by dependent variable Y

$x$ = set of values taken by independent variable X

$\beta_0$ = y intercept

$\beta_1$ = slope

$\boldsymbol{\varepsilon}$ = random error component

# Linear regression line

In context with our example,

$$\text{Premium} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{ Mileage} + \boldsymbol{\varepsilon}$$

y = set of values taken by dependent variable, Premium

x = set of values taken by independent variable, Mileage

$\beta_0$ = premium value where the best fit line cuts the Y - axis (Premium)

$\beta_1$ = beta coefficient for Mileage

ε = random error component

| Mileage | Premium (in dollars) |
|---|---|
| 15 | 392.5 |
| 14 | 46.2 |
| 17 | 15.7 |
| 7 | 422.2 |
| 10 | 119.4 |
| 7 | 170.9 |
| 20 | 56.9 |
| 21 | 77.5 |
| 18 | 214 |
| 11 | 65.3 |
| 7.9 | 250 |
| 8.6 | 220 |
| 12.3 | 217.5 |
| 17.1 | 140.88 |
| 19.4 | 97.25 |

# What is the error term?

In context with our example,

$$\text{Premium} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{ Mileage} + \boldsymbol{\varepsilon}$$

y = set of values taken by dependent variable, Premium

x = set of values taken by independent variable, Mileage

$\beta_0$ = premium value where the best fit line cuts the Y - axis (Premium)

$\beta_1$ = beta coefficient for Mileage

$\varepsilon$ = random error component

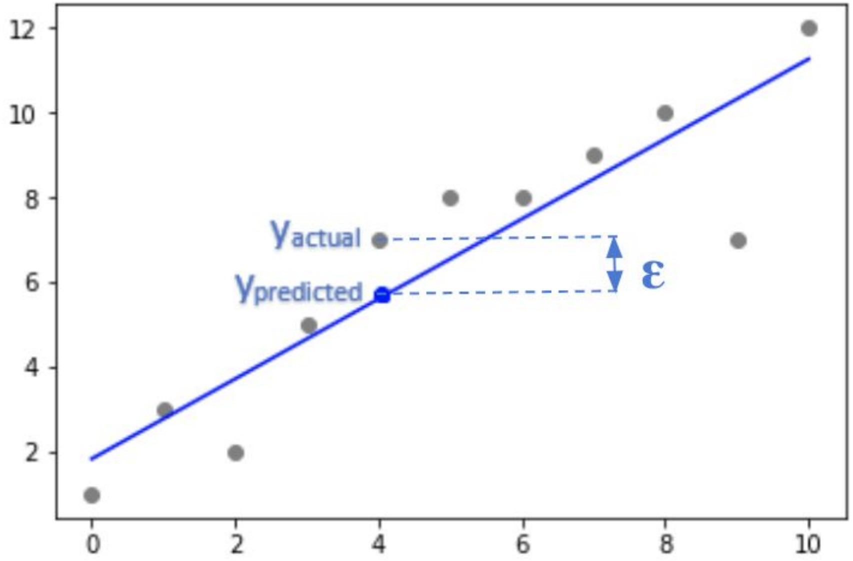
- Error term also called residual represents the distance of the observed value from the value predicted by regression line

- In our example,

  Error term = Actual Premium - Predicted Premium

  for each observation

# Calculating the error term




Equation of regression line is given by,

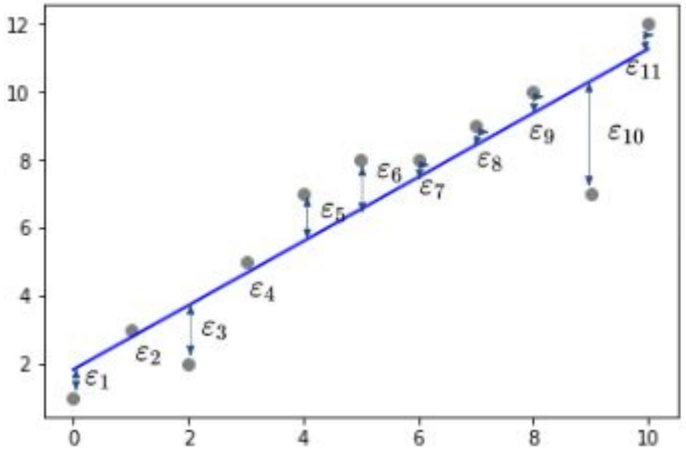
$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x + \boldsymbol{\varepsilon}$$

$$\therefore \boldsymbol{\varepsilon} = y - (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x)$$

$$\therefore \boldsymbol{\varepsilon} = y_{actual} - y_{predicted}$$

# Error calculation

We have an error term for every observation in the data.




We have

$\boldsymbol{\varepsilon}_i = y_{actual} - y_{predicted}$

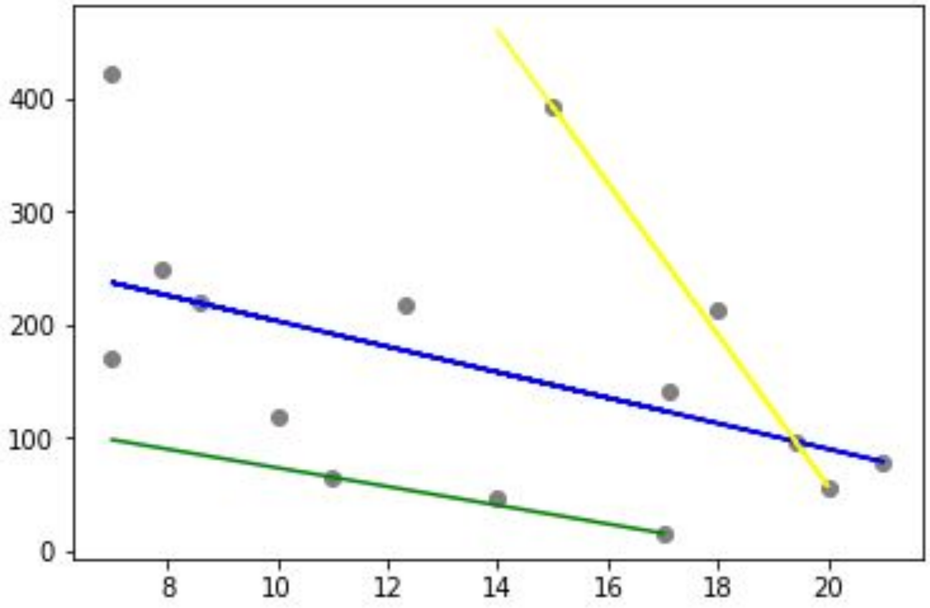
Squared error :

$\boldsymbol{\varepsilon}_i^2 = (y_{actual} - y_{predicted})^2$

Sum of squared errors = $\sum \boldsymbol{\varepsilon}_i^2$

# Ordinary Least Squares Method

# Which line best fits our data?




- The regression line which best explains the trend in the data is the best fit line
- It may pass through all of the points, some of the points or none of the points

# How to obtain the best fit line?

- The ordinary least square method is used to find the best fit line for given data
- This method aims at minimizing the sum of squares of the error terms, that is, it determines those values of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ at which the error terms are minimum

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

# Simple linear regression model

Based on the data and the formulae obtained, the **β** parameters are:

$\boldsymbol{\beta}_0$ =327.0860 and $\boldsymbol{\beta}_1$ = -11.6905.

Thus the model is

$$Y = 327.0860 - 11.6905\ X$$

That is,

$$\text{Premium} = 327.0860 - 11.6905\ \text{Mileage}$$

| Mileage | Premium (in dollars) |
|---|---|
| 15 | 392.5 |
| 14 | 46.2 |
| 17 | 15.7 |
| 7 | 422.2 |
| 10 | 119.4 |
| 7 | 170.9 |
| 20 | 56.9 |
| 21 | 77.5 |
| 18 | 214 |
| 11 | 65.3 |
| 7.9 | 250 |
| 8.6 | 220 |
| 12.3 | 217.5 |
| 17.1 | 140.88 |
| 19.4 | 97.25 |

# Interpretation of β coefficients

- $\beta_1$ gives the amount of change in response variable per unit change in predictor variable
- $\beta_0$ is the y intercept which means when X=0, Y is $\beta_0$
- β's  have an associated p value, which is used to assess its significance in prediction of response variable
- Depending on whether  β's take a positive value k or - k the response variable increases or decreases respectively by k units for every one unit increment in a predictor variable, keeping all other predictor variables constant

# Interpreting the β coefficients

In context with our example,

- $\boldsymbol{\beta}_0$ = 327.0860: represents the premium of a car immediately after manufacture (i.e. Mileage = 0)

- $\boldsymbol{\beta}_1$ = - 11.6905: is the average decrease in the premium of a car due to unit increase in mileage.

Note: For mileage = 0, the premium is equal to $\beta_0$ = $ 327.0860.

# How is the $y_{predicted}$ obtained?

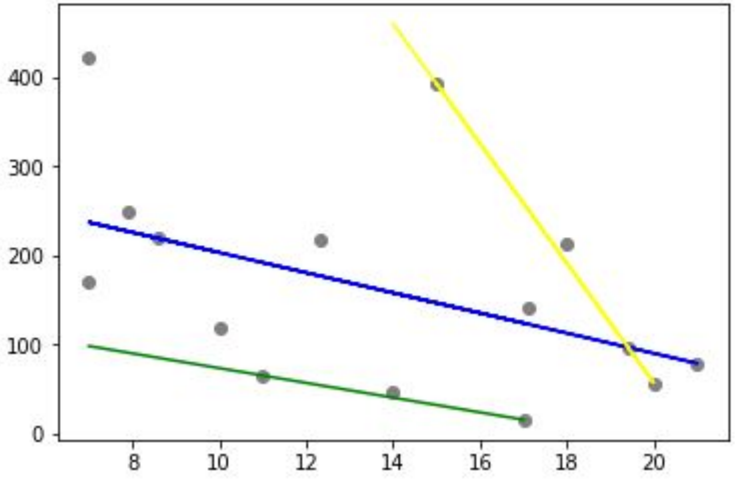
Substitute the values for X in the model.

For example:

For mileage (x) = 17, the predicted premium, ($y_{predicted}$) is obtained as

$$y_{predicted} = 327.0860 - 11.6905 * 17 = \$ \ 128.3475$$

# Simple regression - best fit line




| $\sum \varepsilon^2$ | $\sum \varepsilon^2$ | $\sum \varepsilon^2$ |
|---|---|---|
| $3.94 \times 10^5$ | $1.6 \times 10^5$ (Least Error) | $26.8 \times 10^5$ |

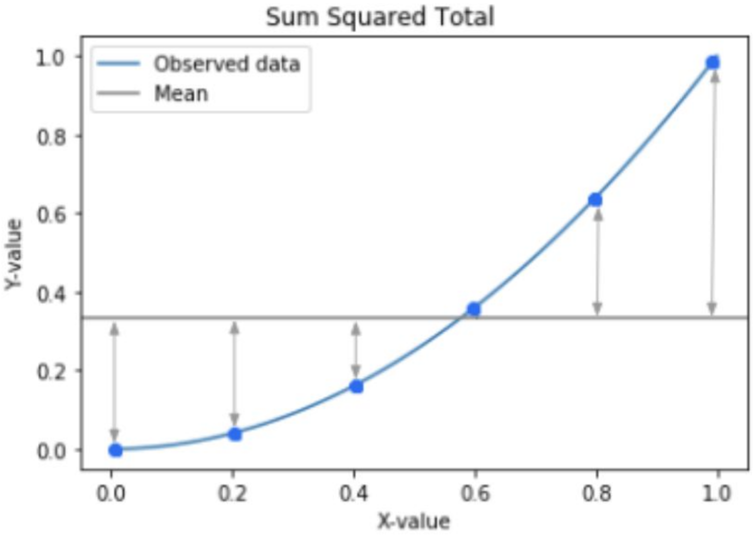
Since the blue line has least error it is the best fit line

# Interview Questions

- Explain the math behind Linear Regression.
- Difference between regression and ANOVA.
- What is a residual?
- What does the residual plot look like?
- What is linear regression?
- What is F value?

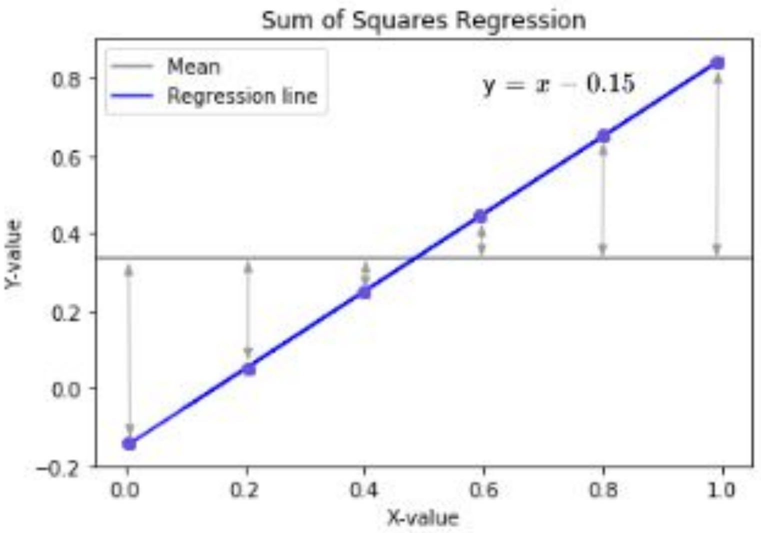
# Measures of Variation

# Sum of squares total



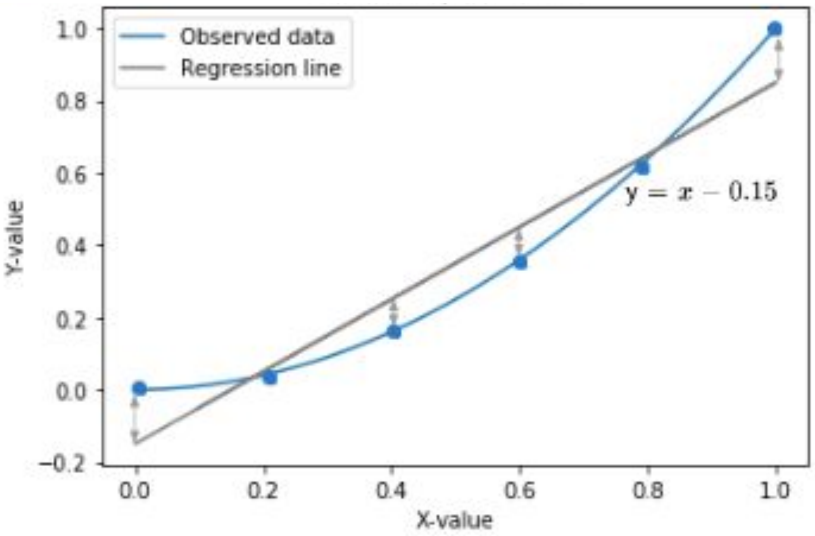

- The sum of squares total (SST) is the sum of squared differences between the observed response variable and its mean
- It can be seen as the total variation of the response variable about its mean value
- SST is the measure of variability in the response variable without considering the effect of predictor variables
- Also known as Total Sum of Square (TSS)

# Sum of squares regression




- The sum of squares regression (SSR) is the sum of squared differences between the predicted value and the mean of the response variable
- SSR is the measure of variability in the response variable considering the effect of predictor variable . It is the explained variation
- It is the explained variation
- Also known as Regression Sum of Square (RSS)

# Sum of squares of error



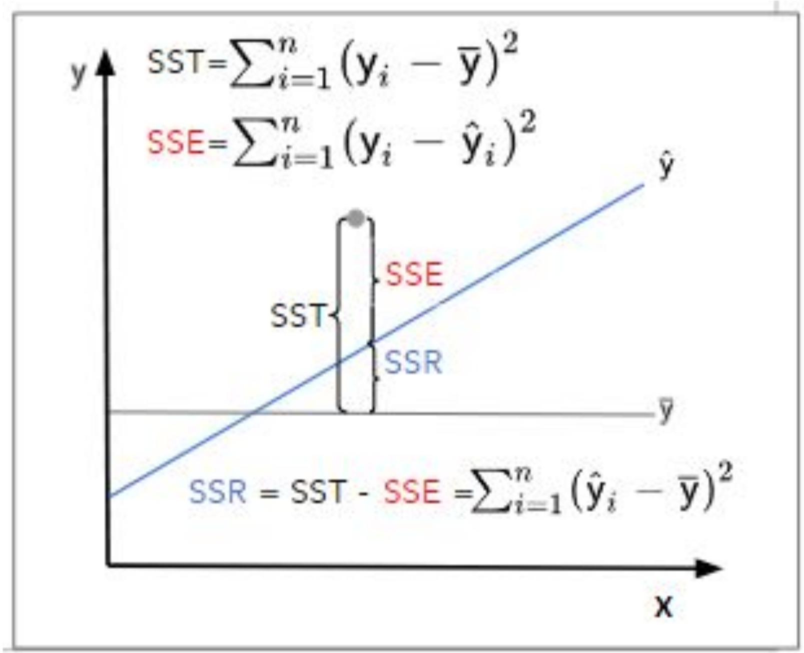

- The sum of squares of error (SSE) is the sum of squared differences between observed response variable and its predicted value

- SSE is the measure of variability in the response variable remaining after considering the effect of predictor variables

- It is the unexplained variation

- Also known as Error Sum of Square (ESS)

# Variation in response variable




$y_i$ = observed values of y

$\hat{y}_i$ = predicted values of y

$\bar{y}$ = mean value of variable y

# Total variation

Total variation = Explained variation + Unexplained variation

$$SST = SSR + SSE$$

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

# Measure of unexplained variation

- Standard error of estimate is a measure of the unexplained variance
- Smaller value of standard error of estimate indicates a better model

$$Sxy = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k}}$$

n = sample size

k = number of parameter estimates ($\beta_0, \beta_1$)

# Measure of explained variation

$R^2$ also called the coefficient of determination gives total percentage of variation in Y that is explained by predictor variable.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{SSR}}{\text{SST}} \qquad 0 \leq R^2 \leq 1$$

$$R^2 = 1 - \frac{SSE}{SST}$$

# R-squared

- Since $0 \leq SSE \leq SST$, mathematically we have $0 \leq R^2 \leq 1$
- $R^2$ assumes that all the independent variables explain the variation in dependent variable
- For simple linear regression, the squared correlation between the response variable Y and independent variable X is the $R^2$ value
- For our model, $R^2 = 0.226$. It implies that 22.6% variation in premium amounts is explained by the mileage of a car

# Demerits of R-squared

- The value of $R^2$ increases as new numeric predictors are added to the model, it may appear that it is a better model, which can be misleading

- Also, if the model has too many variables, the model is feared to be overfitted. Overfitted data generally has a high $R^2$ value.

# Multiple Linear Regression

# Multiple linear regression

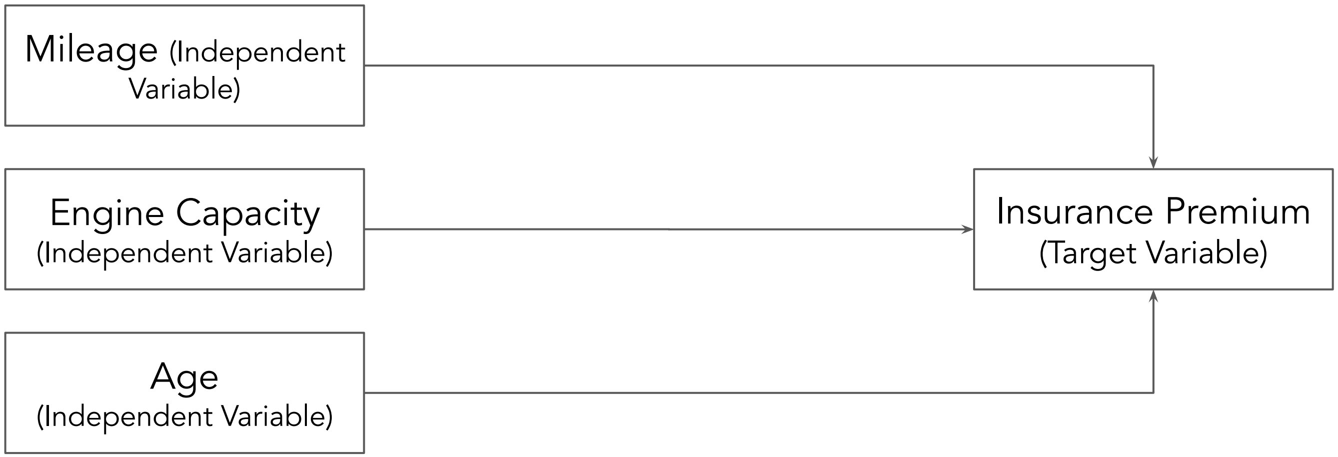
Multiple regression model is used when multiple predictor variables $[X_1, X_2, X_3, \ldots, X_n]$ are used to predict the response variable Y

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n + \varepsilon$$

$\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \ldots, \boldsymbol{\beta}_n$ are the parameters of the linear regression model with n independent variables

# Variable that contributes to Insurance Premium

Let us consider impact of a multiple variables on the Insurance Premium




We say that only Mileage, Engine Capacity and Age decide what the insurance premium should be.

# Data

Let us consider the following data.

| Mileage | Engine_Capacity | Age | Premium (in dollars) |
|---|---|---|---|
| 15 | 1.8 | 2 | 392.5 |
| 14 | 1.2 | 10 | 46.2 |
| 17 | 1.2 | 8 | 15.7 |
| 7 | 1.8 | 3 | 422.2 |
| 10 | 1.6 | 4 | 119.4 |
| 7 | 1.4 | 3 | 170.9 |
| 20 | 1.2 | 7 | 56.9 |
| 21 | 1.6 | 6 | 77.5 |
| 18 | 1.2 | 2 | 214 |
| 11 | 1.6 | 5 | 65.3 |
| 7.9 | 1.4 | 3 | 250 |
| 8.6 | 1.6 | 3 | 220 |
| 12.3 | 1.2 | 2 | 217.5 |
| 17.1 | 1.6 | 1 | 140.88 |
| 19.4 | 1.2 | 6 | 97.25 |

# Multiple Linear Regression equation

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \boldsymbol{\beta}_3 x_3 + \ldots + \boldsymbol{\beta}_n x_n + \boldsymbol{\varepsilon}$$

$y$ = set of values taken by dependent variable Y

$x_i$ = set of values taken by independent variable $X_i$ , $i \in [1,n]$

$\beta_0$ = y intercept

$\beta_i$ = beta coefficient for the $i^{th}$ independent variable $X_i$, $i \in [1,n]$

$\boldsymbol{\varepsilon}$ = random error component

# Linear regression for our example

$$\text{Premium} = \beta_0 + \beta_1 \text{ Mileage} + \beta_2 \text{ Engine\_Capacity} + \beta_3 \text{Age} + \varepsilon$$

| | Description |
|---|---|
| Premium | Set of values taken by the variable Premium |
| $\beta_0$ | Premium value where the best fit line cuts the Y-axis (Premium) |
| $\beta_1$ | Regression coefficient of variable Mileage |
| Mileage | Set of values taken by the variable Mileage |
| $\beta_2$ | Regression coefficient of variable Engine_Capacity |
| Engine_Capacity | Set of values taken by the variable Engine_Capacity |
| $\beta_3$ | Regression coefficient of variable Age |
| Age | Set of values taken by the variable Age |
| $\varepsilon$ | Error component |

# Multiple linear regression model

Based on the data and the formulae obtained, the β parameters are:

$\beta_0 = 138.398$, $\beta_1 = - 4.876$,

$\beta_2 = 137.633$ and $\beta_3 = - 23.718$.

| Mileage | Engine_Capacity | Age | Premium (in dollars) |
|---|---|---|---|
| 15 | 1.8 | 5 | 392.5 |
| 14 | 1.2 | 5 | 46.2 |
| 17 | 1.2 | 5 | 15.7 |
| 7 | 1.8 | 10 | 422.2 |
| 10 | 1.6 | 4 | 119.4 |
| 7 | 1.4 | 5 | 170.9 |
| 20 | 1.2 | 3 | 56.9 |
| 21 | 1.6 | 4 | 77.5 |
| 18 | 1.2 | 4 | 214 |
| 11 | 1.6 | 5 | 65.3 |
| 7.9 | 1.4 | 3 | 250 |
| 8.6 | 1.6 | 5 | 220 |
| 12.3 | 1.2 | 2 | 217.5 |
| 17.1 | 1.6 | 6 | 140.88 |
| 19.4 | 1.2 | 2 | 97.25 |

Thus the model is

$$Y = 138.398 - 4.876 x_1 + 137.633 x_2 - 23.718 x_3$$

That is,

Premium = 138.398 - 4.876 Mileage + 137.633 Engine_Capacity - 23.718 Age

# Interpreting the β coefficients

In context with our example,

- $\beta_0$ = 138.398: the value of premium when the mileage, engine capacity and age are all equal to 0 (which is absurd)
- $\beta_1$ = - 4.876: is the average decrease in the premium of cars due to unit increase in mileage, all else held constant.
- $\beta_2$ = 137.633: the average increase in the premium of the cars due to engine capacity, all else held constant.
- $\beta_3$ = - 23.718: the average decrease in the premium of the cars due to age, all else held constant.

# Revisiting R-squared

$R^2$ also called the coefficient of determination gives total percentage of variation in Y that is explained by predictor variable.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{SSR}}{\text{SST}} \qquad 0 \leq R^2 \leq 1$$

$$R^2 = 1 - \frac{SSE}{SST}$$

# Adjusted R-squared

Adjusted $R^2$ gives the percentage of variation explained by independent variables that actually affect the dependent variable

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

$R^2$ = R squared value for model

n = sample size

k = number of features

# Adjusted R-squared

- $R^2_{adj} \leq R^2$ (always)

- As the number of independent variables in the model increase, the adjusted $R^2$ will decrease unless the model significantly increases the $R^2$

- So to know whether addition of a variable explains the variation of the response variable, compare the $R^2_{adj}$ values along with $R^2$

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

As k (no. of independent variables) increases, value of (n-k-1) decreases

# Interview Question

- What is the difference between r-squared and adjusted r-squared?
- What is R-squared and where is it used?

# Thank You