Introduction to Machine Learning

Supervised Linear Regression

Agenda

- Machine Learning Overview
- Traditional Programming Vs Machine Learning
- Understanding the Problem and Data
- Steps in Machine Learning
- Basic Terms used in Machine Learning
- Types of Machine Learning
- Applications of machine learning: Use Cases
- Measures of dispersion and Central Tendency

Agenda

- Simple Linear Regression
- Regression Analysis
- Ordinary Least squares Method
- Measures of Variation
- Inferences about slope
- Multiple Linear Regression

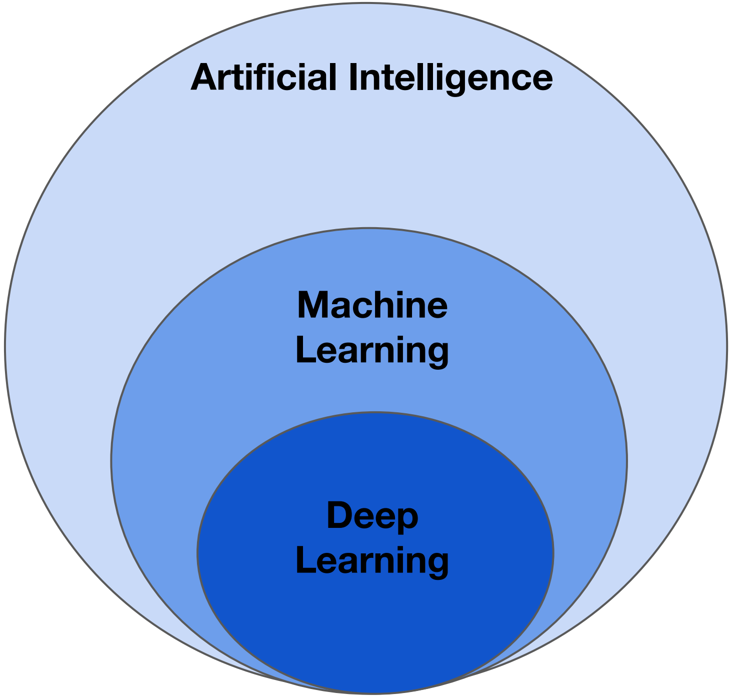
Machine Learning Overview

- Machine Learning is the science to make computers learn from data without programming them explicitly and improve their learning over time in an autonomous fashion.

- This learning comes by feeding the data in the form of observations and real-world interactions.

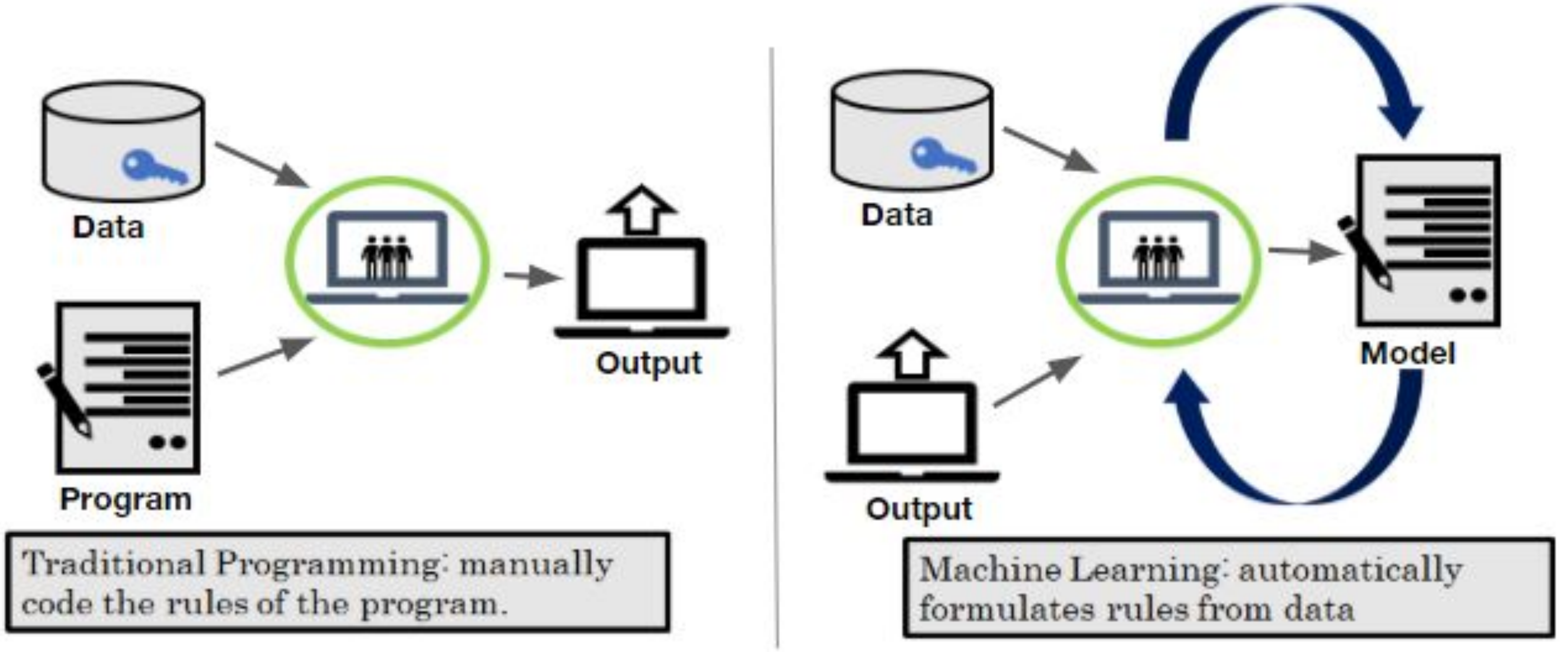
- Machine Learning can also be defined as a tool to predict future events or values using past data.

AI Vs ML Vs DL



- Artificial Intelligence: Infusing intelligence in machines
- Machine Learning: Algorithms that “learn” from experience/data
- Deep Learning: Algorithms inspired by human brain, that can ***learn features*** from large data

Traditional Programming vs. Machine Learning



Traditional Programming: manually code the rules of the program.

Machine Learning: automatically formulates rules from data

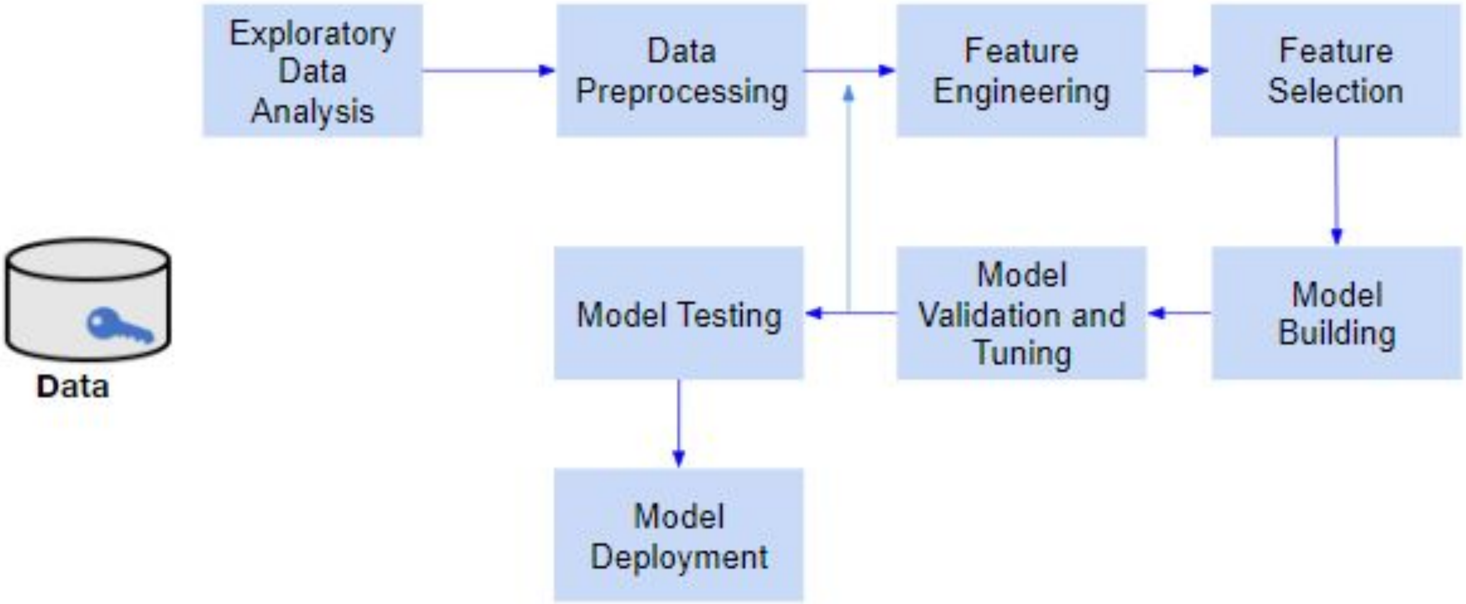
Understanding the Problem Statement

- What is the domain and context?
- What business problem are you trying to solve?
- What is the return on investment ?
- Does this solution require machine learning?
- If machine learning is required, what type of ML task is it?
- What is the suitable evaluation metric for this?

Data Collection

- Manual data collection / Using available data
- Collecting data from multiple sources with the help of data engineers and business
- Merging and joining different datasets as required to solve the problem.
- Maintaining version of dataset for future reference
- If data is too big, take a subset of data to work with.

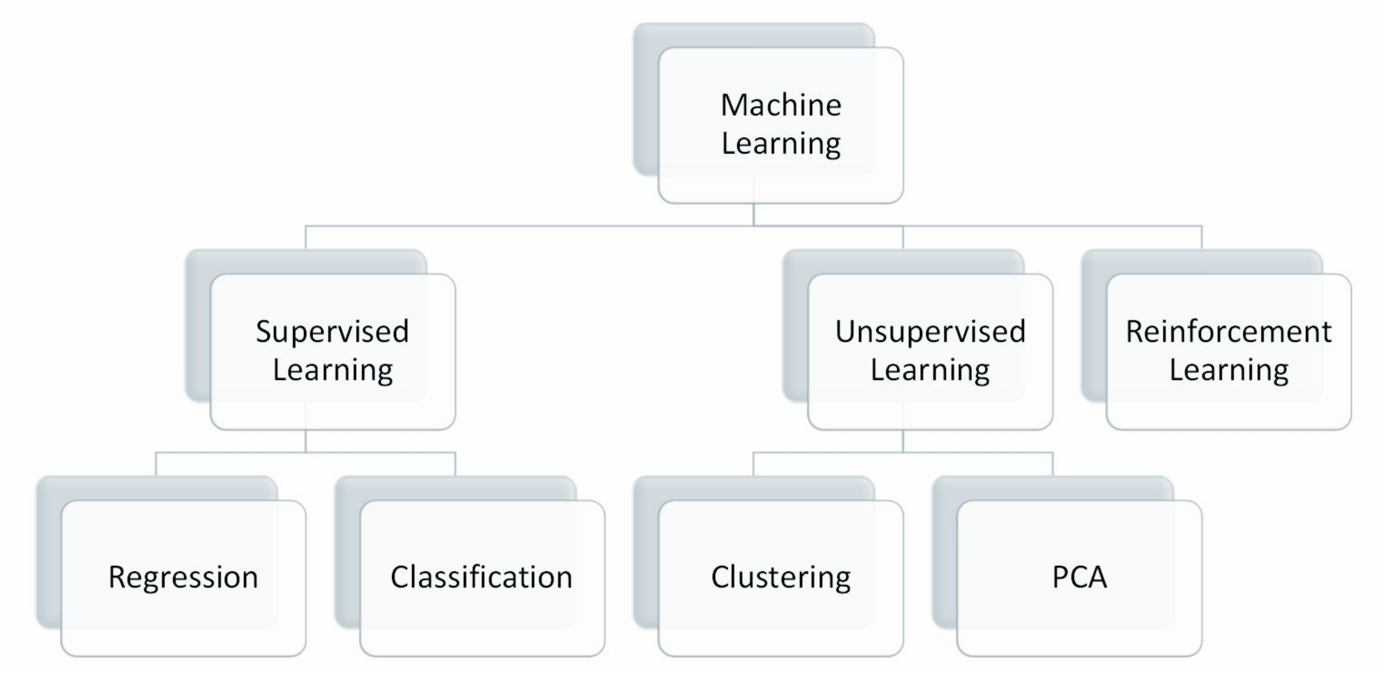
Steps in Machine Learning Algorithm



Types of Machine Learning

- Supervised Learning - Training happens based on labelled data
- Unsupervised Learning - Meant to recognise patterns in unlabelled data
- Reinforcement Learning - Machine gets rewarded for right outcome

Types of Machine Learning



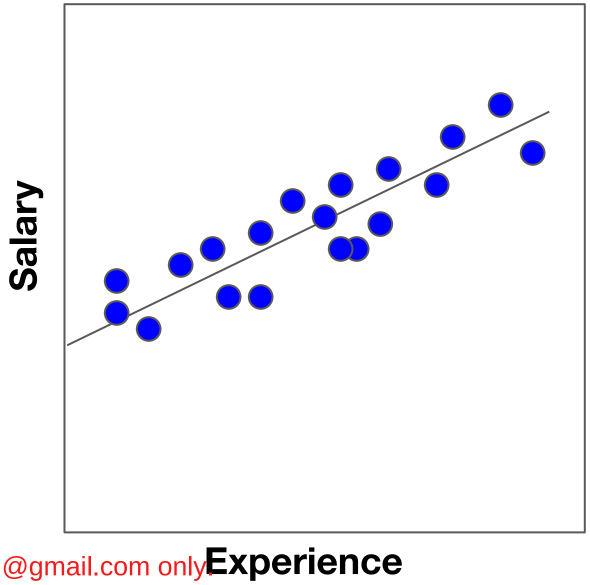
Supervised learning

- Class of machine learning that work on externally supplied instances in form of predictor attributes and **associated target values**.
- The model learns from the training data using these **‘target variables’** as reference variables.
    - Ex1 : model to predict the resale value of a car based on its mileage, age, color etc.
- The **target values** are the ‘correct answers’ for the predictor model which can either be a **regression model** or a **classification model**.

Supervised learning- Regression

- Linear Regression
- kNN regressor
- SVR
- Decision tree regressor
- Random forest regressor
- Neural Networks

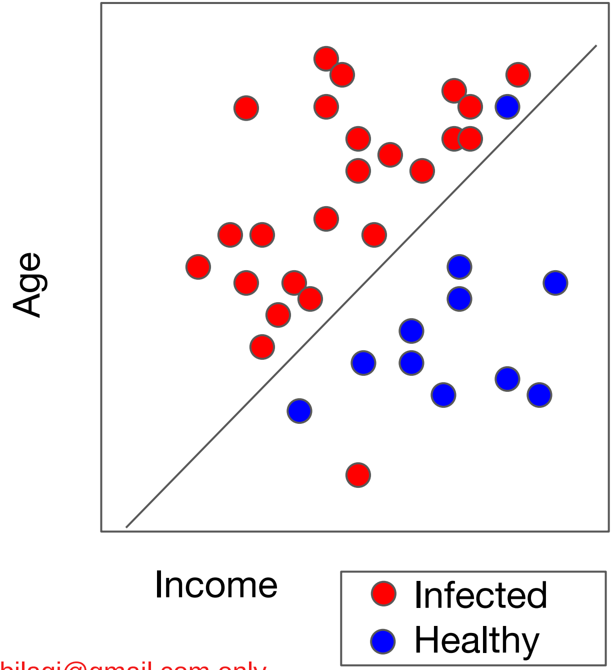
Predicting Salary from Experience in a Profession (say Teaching)



Supervised learning- Classification

- Logistic regression
- k Nearest Neighbours
- Decision tree
- Support vector machines
- Random forest
- Naive bayes

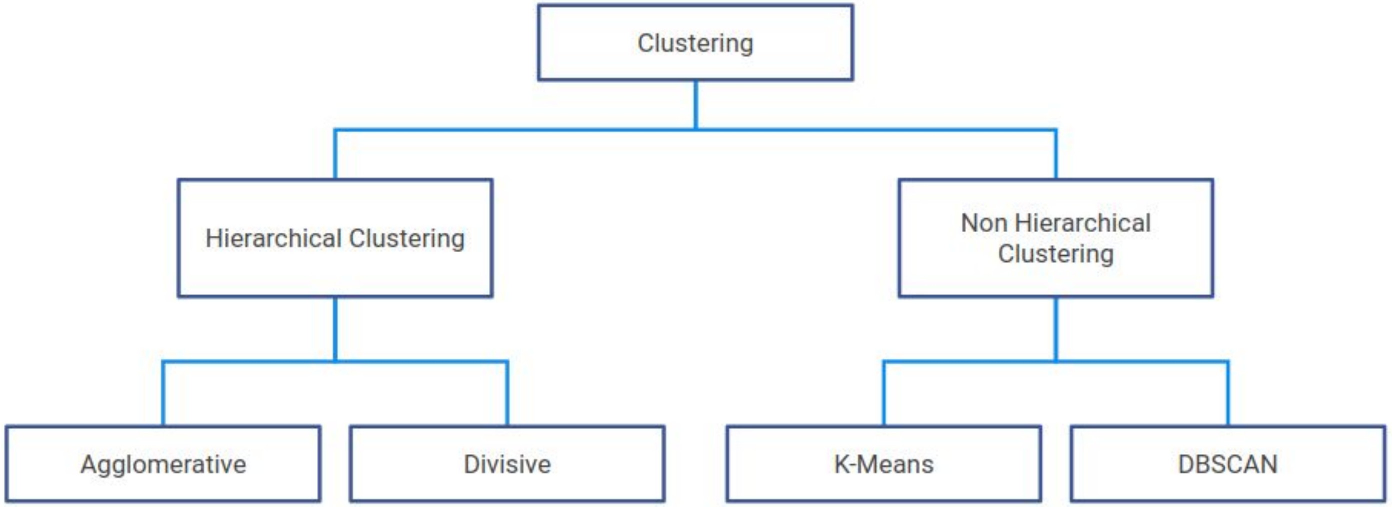
Predicting whether a person is healthy or Infected



Unsupervised learning

- K-means clustering
- Hierarchical clustering
- Principal component analysis
- Hidden Markov Model
- FP-Growth
- Apriori Analysis

Unsupervised Learning- Clustering



Machine Learning Prerequisites

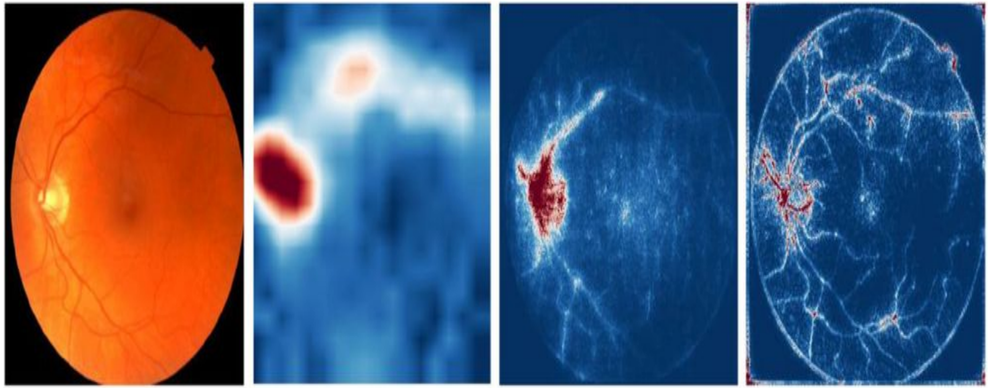
For the practical Machine Learning that we are going to be dealing with in our course, we will require a decent understanding of

- Linear Algebra
- Calculus
- Statistics
- Programming

Nevertheless, these prerequisites are not rigid but flexible in keeping with what we want to achieve. From designing a new algorithm to dragging and dropping ML objects to aid in running a business.

Use cases: Detecting Diseases from X-rays/Images

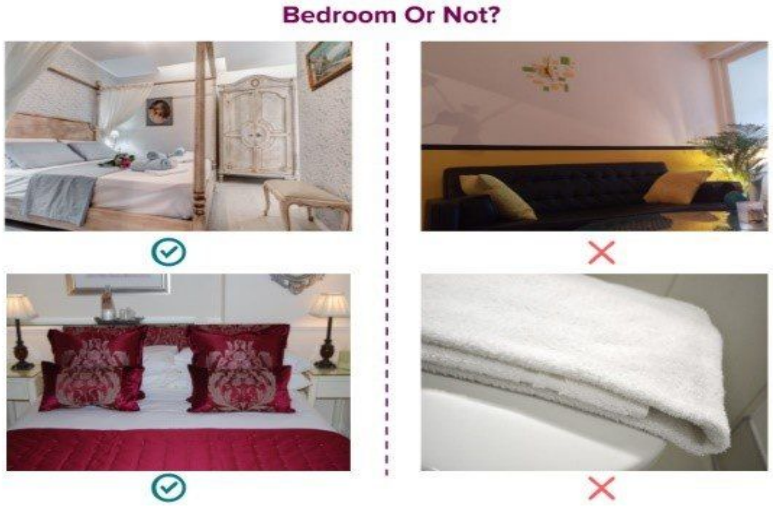
- Anemia is a major health problem that causes dizziness,weakness & tiredness.
- Deep learning model can quantify hemoglobin using images of the back of the eye and other data such as age,gender.
- Easier to use than blood test & Non-destructive testing



Courtesy: https://blog.google/technology/health/anemia-detection-retina/

Use cases: Pricing & Customer Satisfaction

- Airbnb in hospitality industry face issues in personalization, pricing & improving the guest experience
- Uses ML to personalize search rankings for guests, optimizes pricing for hosts .
- Natural language processing to understand guest reviews.
- Uses image classification to improve search rankings by photos based on what guests care about the most .



Every time you interact with an Airbnb app or the website, you're interacting with machine learning in some way or another."

– Mike Curtis, VP of Engineering, Airbnb

Courtsey:https://digital.hbs.edu/platform-rctom/submission/airbnb-utilizing-machine-learning-to-optimize-travel/

Use Cases: Personalized Recommendations

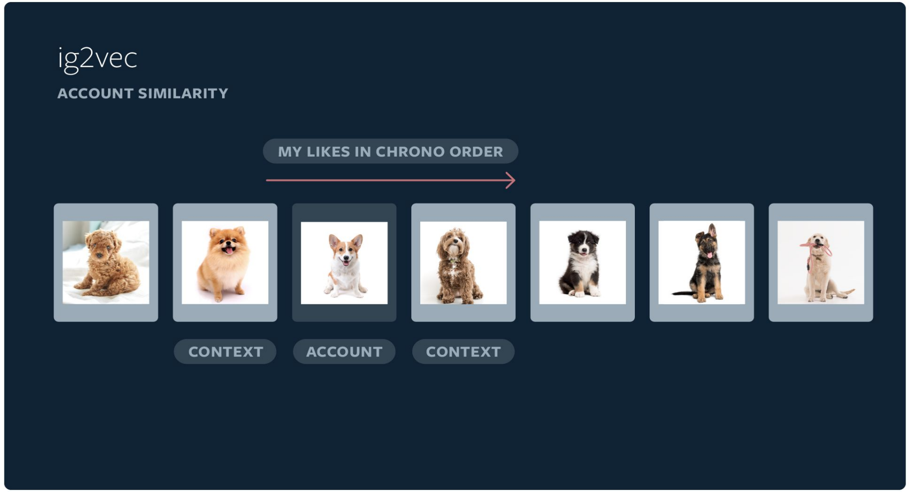
- Ecommerce firms such as Amazon, Flipkart faces a challenge of understanding each customer and what to recommend to each person

- The firms considers all the purchases made by the said user and also studies the behaviour of multiple users and their buying/consumption behaviour and comes up with an automated recommendation algorithm based on user and item

- They are providing personalized recommendations without spending time and effort on each user as done by any traditional seller

Use Cases: AI for Instagram Recommendations

- Over half of the Insta community visits Instagram Explore every month to discover new photos, videos, and Stories.
- Recommending the most relevant content out of billions causes multiple ML challenges.
- An algorithm identifies long-term interests
- Another algorithm identifies recommendations based on recent content.Face tagging
- different application



Courtesy:https://ai.facebook.com/blog/powered-by-ai-instagrams-explore-recommender-system/

Revisiting Descriptive Statistics

- **Concerned with Data Summarization, Graphs/Charts, and Tables.**
- **Also called as summary statistics**
  - **Measure of central tendency - mean, mode, median**
  - **Measure of statistical dispersion - variance, standard deviation, range**
  - **Measure of shape of a distribution - skewness, kurtosis**
  - **Measure of statistical dependence - Pearson correlation**
- **Common techniques - box plot, histogram**

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.

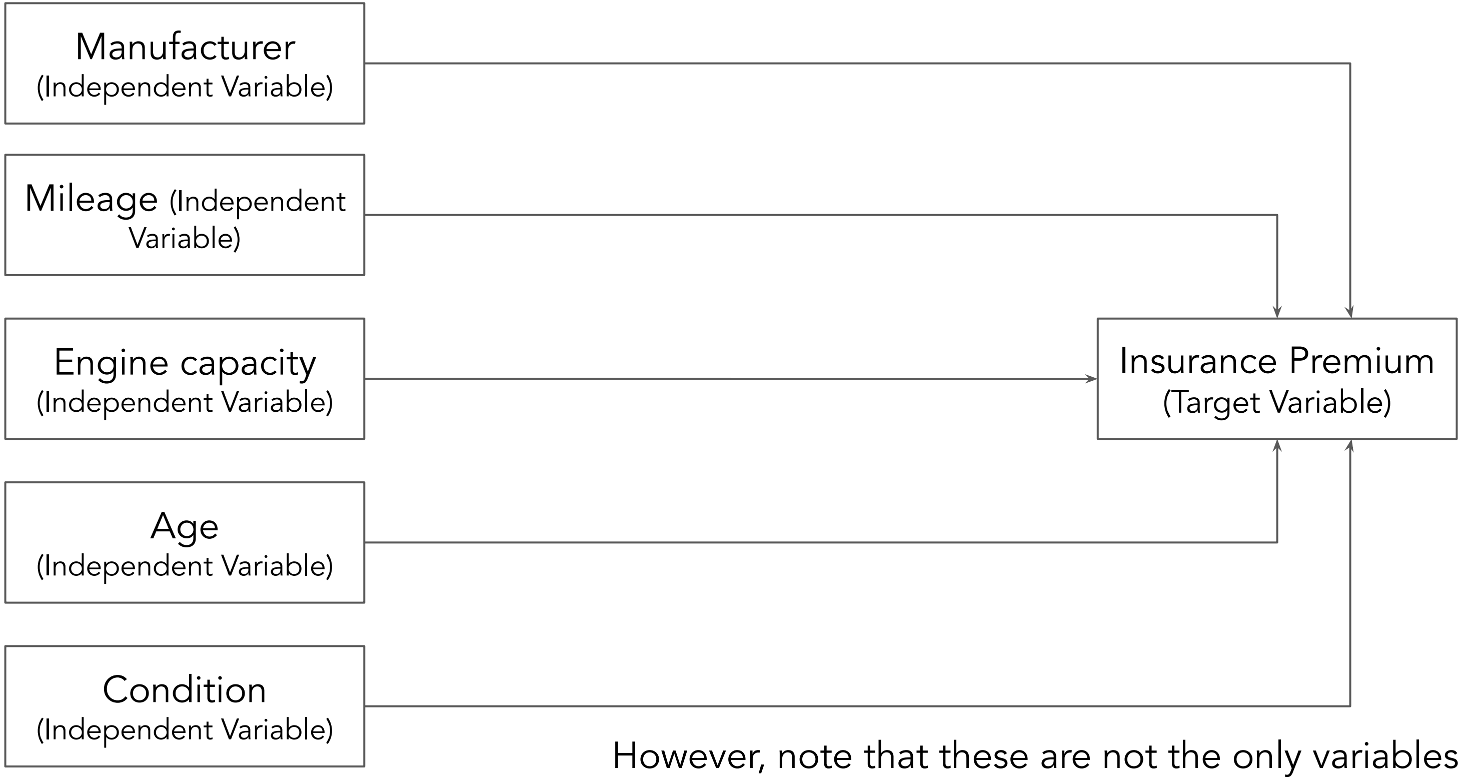
Dependent variable

- 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
- Usually denoted by X
- Independent Variable = Predictor Variable
- In our example, Age, Mileage and Condition of the car are the independent variables

Variables that may contribute to insurance premium



Manufacturer (Independent Variable)

Mileage (Independent Variable)

Engine capacity (Independent Variable)

Age (Independent Variable)

Condition (Independent Variable)

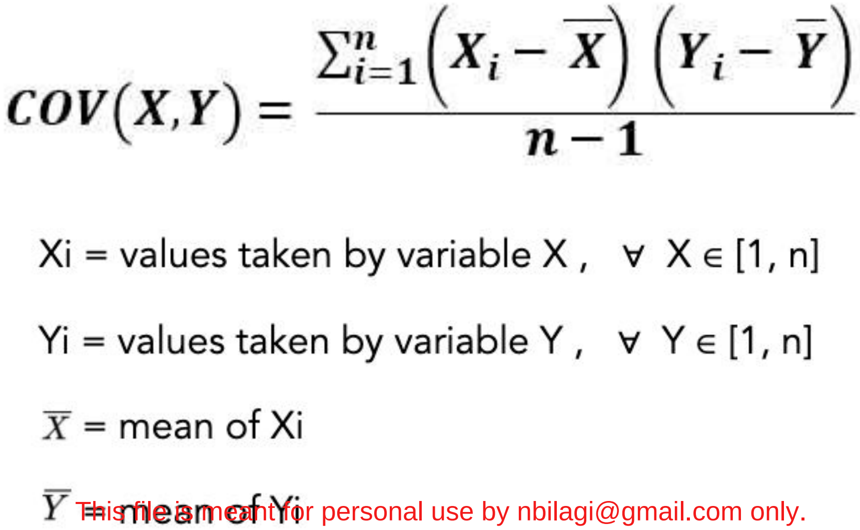
Insurance Premium (Target Variable)

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

Visiting Basics

Covariance

Covariance is a measure of how changes in one variable are associated with changes in anothe



$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}$$

Xi = values taken by variable X , $\forall$ X $\in$ [1, n]

Yi = values taken by variable Y , $\forall$ Y $\in$ [1, n]

$\overline{X}$ = mean of Xi

$\overline{Y}$ = mean of Yi

Pearson’s correlation coefficient

Correlation is a measure for linear association between two numeric variables.

$$R = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

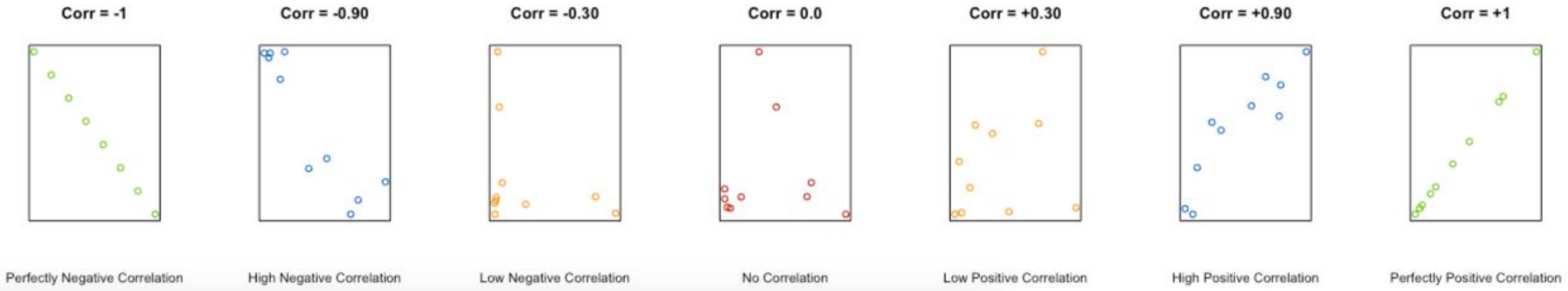
Cov(x, y) = covariance of variables x and y

σ_x = standard deviation of x

σ_y = standard deviation of y

Value of correlation

Correlation is a scaled version of covariance that takes on values in [−1,1] with a correlation of ±1 indicating perfect linear association and 0 indicating no linear relationship.

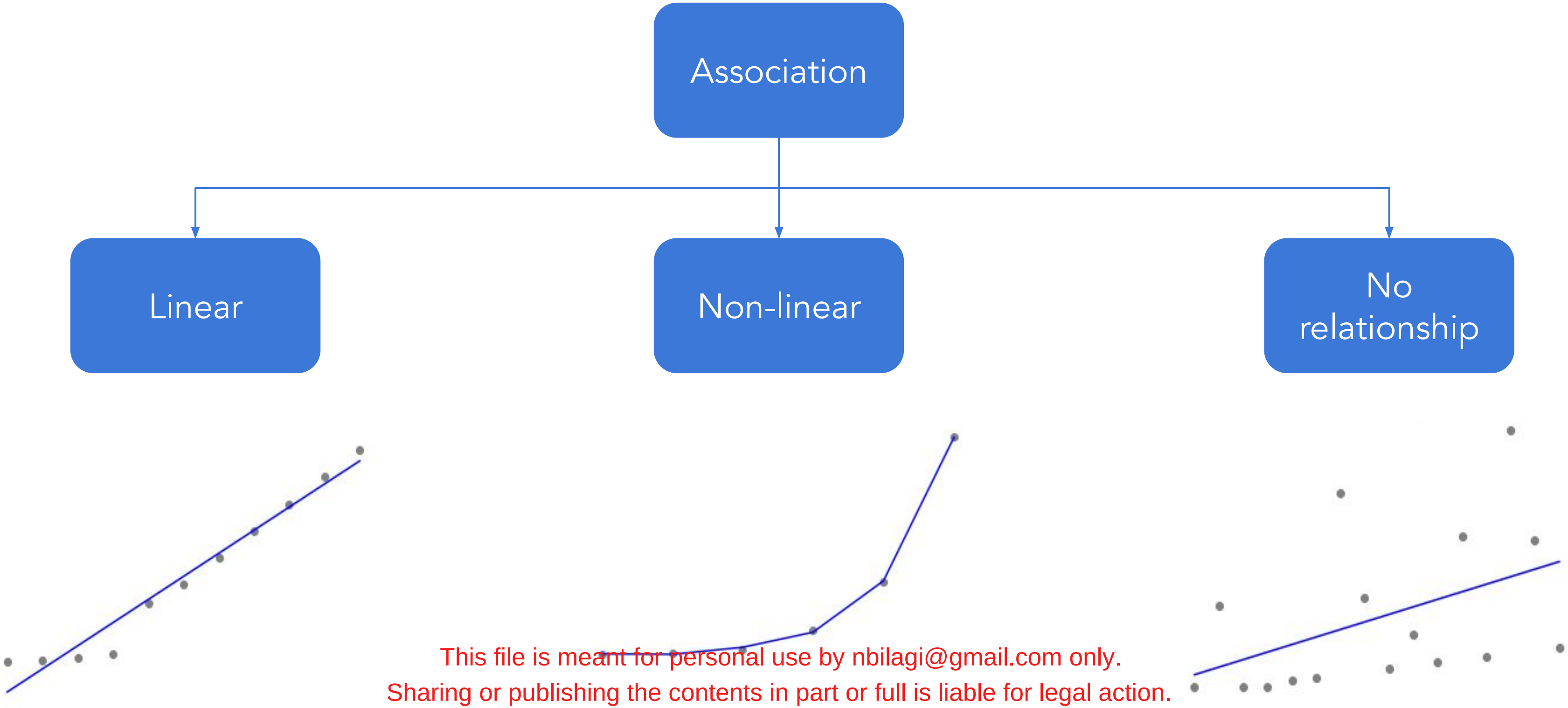


Regression Analysis

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



Association

Linear

Non-linear

No relationship

Simple linear regression

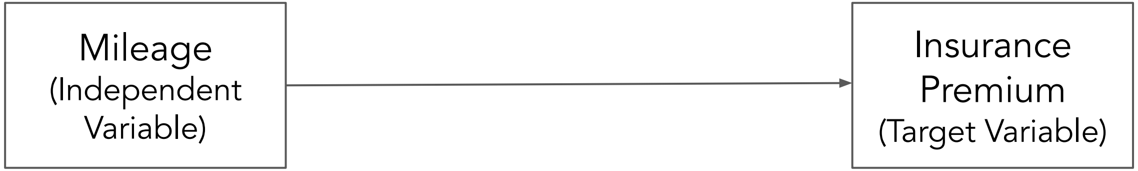
A simple linear regression model (also called **bivariate regression**) has one independent variable X that has a linear relationship with the dependent variable Y

$$\mathbf{y = \beta_0 + \beta_1 x + \varepsilon}$$

$\boldsymbol{\beta_0}$ 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.



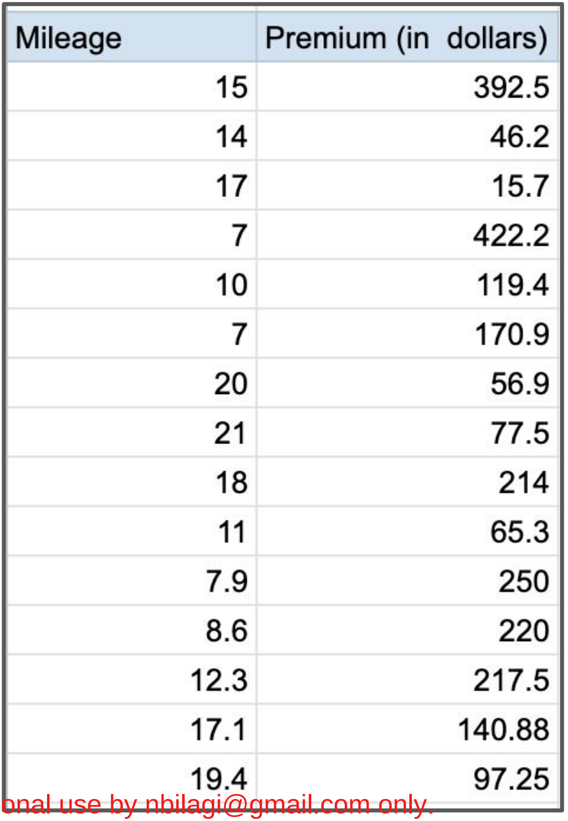
Mileage
(Independent Variable)

Insurance Premium
(Target Variable)

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

$$\mathbf{y = \beta_0 + \beta_1 x + \varepsilon}$$

y = set of values taken by dependent variable Y

x = set of values taken by independent variable X

β_0 = y intercept

β_1 = slope

ε = random error component

Linear regression line

In context with our example,

$$\mathbf{Premium = \beta_0 + \beta_1\, Mileage + \varepsilon}$$

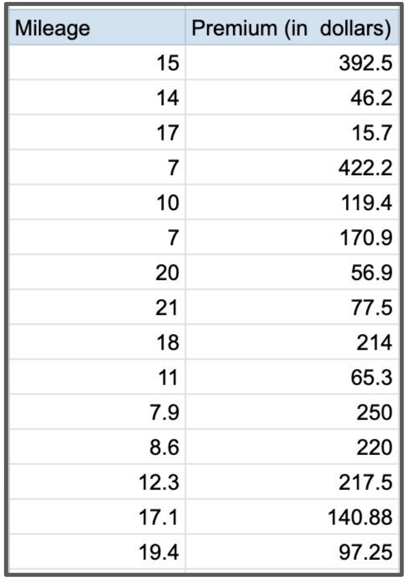
y = set of values taken by dependent variable, Premium

x = set of values taken by independent variable, Mileage

β_0 = premium value where the best fit line cuts the Y - axis (Pre

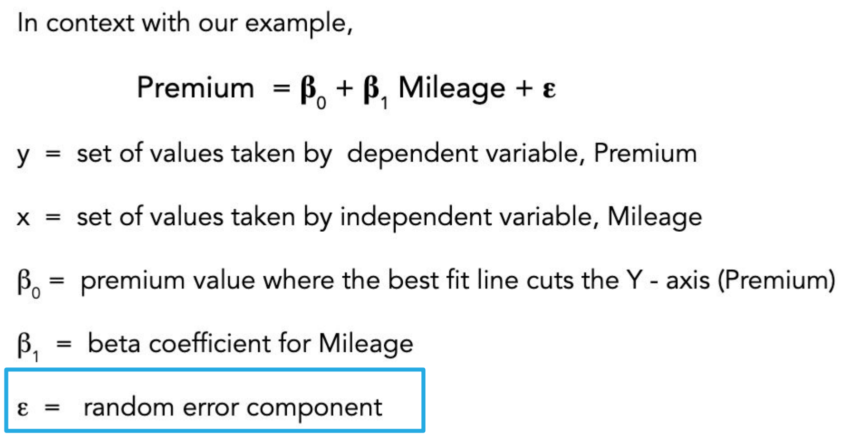
β_1 = beta coefficient for Mileage

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

β_0 = premium value where the best fit line cuts the Y - axis (Premium)

β_1 = beta coefficient for Mileage

ε = random error component

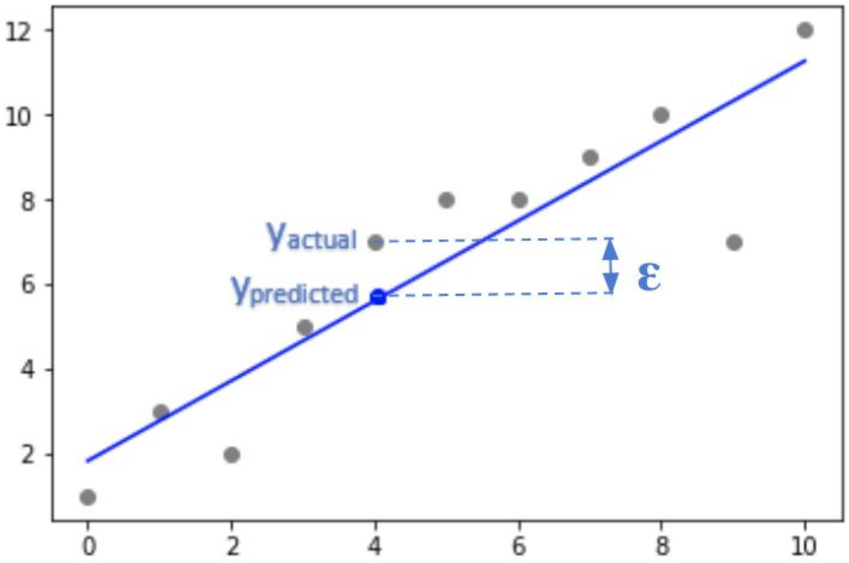
- Error term also called residual represents the distance of the observed value from the value predicted by regression line

- In our example,

  Error term = Actual Premium - Predicted Premium

  for each observation

Calculating the error term



Equation of regression line is given by,

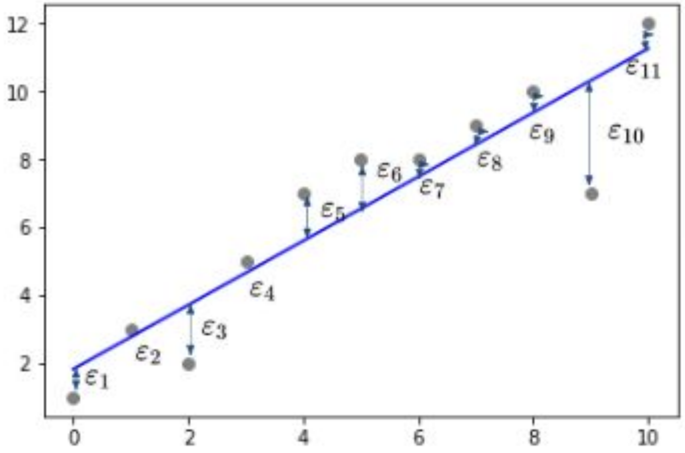
$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\therefore \varepsilon = y - (\beta_0 + \beta_1 x)$$

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

Error calculation

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



We have

$\boldsymbol{\varepsilon_i = y_{actual} - y_{predicted}}$

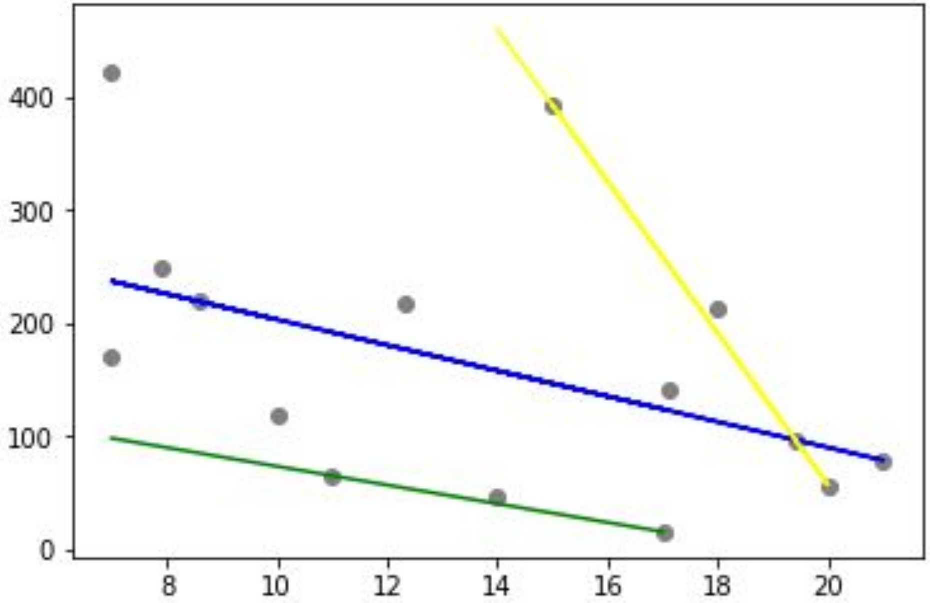
Squared error :

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

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

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 **β_0 and β_1 at which the error terms are minimum**

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

Maths behind OLS

- We have seen that the error term $\boldsymbol{\varepsilon = y - (\beta_0 + \beta_1 x)}$
- The OLS method minimizes E = $\sum\boldsymbol{\varepsilon^2 = \sum(y - (\beta_0 + \beta_1 x))^2}$
- To minimize the error we take partial derivatives with respect to $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ and equate them to zero

$$\delta E/\delta\beta_0 = 0$$

$$\delta E/\delta\beta_1 = 0$$

- So we get two equations with two unknowns, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

Maths behind OLS

- So we get:

$$\delta E/\delta \boldsymbol{\beta}_0 = \sum 2\,(\mathbf{y} - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_1\mathbf{x})\,(-1) = 0$$

$$\delta E/\delta \boldsymbol{\beta}_1 = \sum 2\,(\mathbf{y} - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_1\mathbf{x}))\,(-x_1) = 0$$

- Expanding these equations, we get $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ as:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad \beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

Simple linear regression model

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

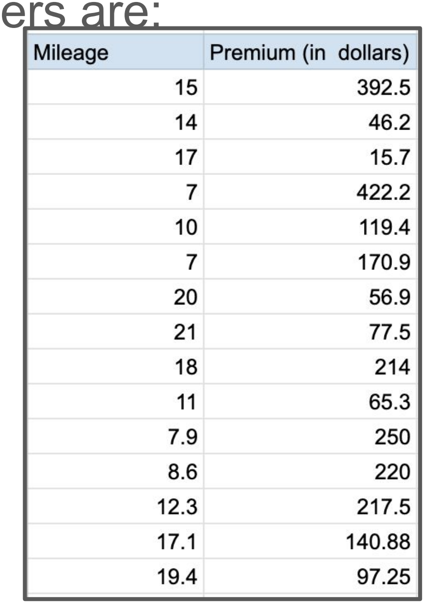
β_0 =327.0860 and β_1 = -11.6905.

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

- β_1 gives the amount of change in response variable per unit change in predictor variable
- β_0 is the y intercept which means when X=0, Y is β_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,

- **β_0 = 327.0860**: represents the premium of a car immediately after manufacture (i.e. Mileage = 0)

- **β_1 = - 11.6905**: the average decrease in the premium of the cars due to the mileage

Note: For mileage = 0, the premium is equal to β_0 = $ 327.0860.

How is the $y_{predicted}$ obtained?

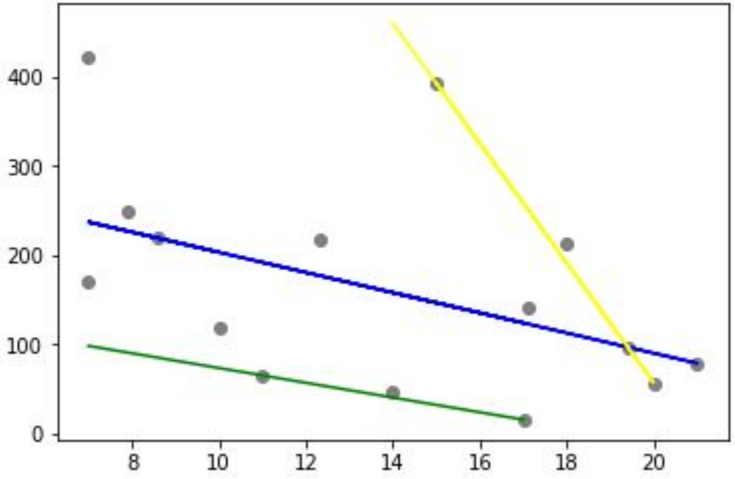
Substitute the values for X in the model.

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

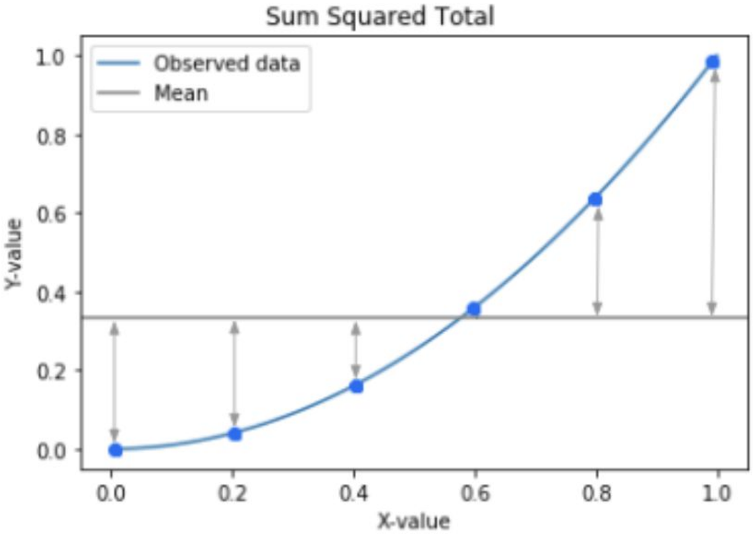


$\sum \varepsilon^2$	$\sum \varepsilon^2$	$\sum \varepsilon^2$
3.94×10^5	1.6×10^5 (Least Error)	26.8×10^5

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

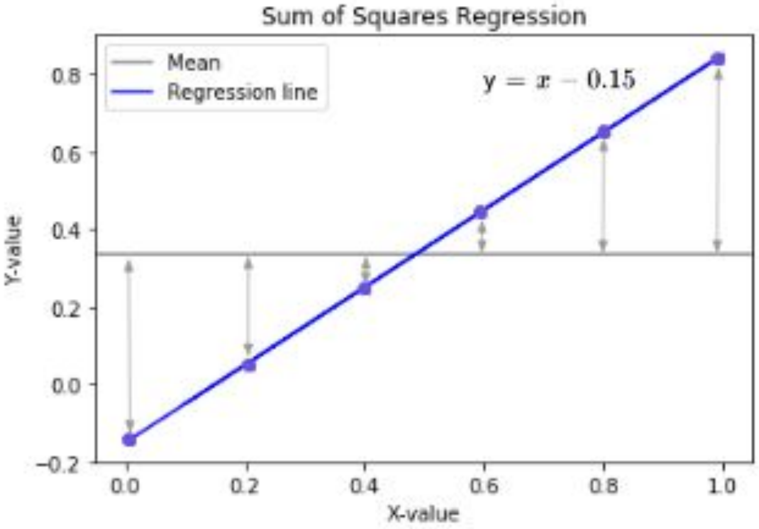
Measures of Variation

Sum of squares total



- The sum of squares total (SST) is the sum of squared differences between the observation 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 dependent variable
- Also known as Total Sum of Square (TSS)

Sum of squares regression



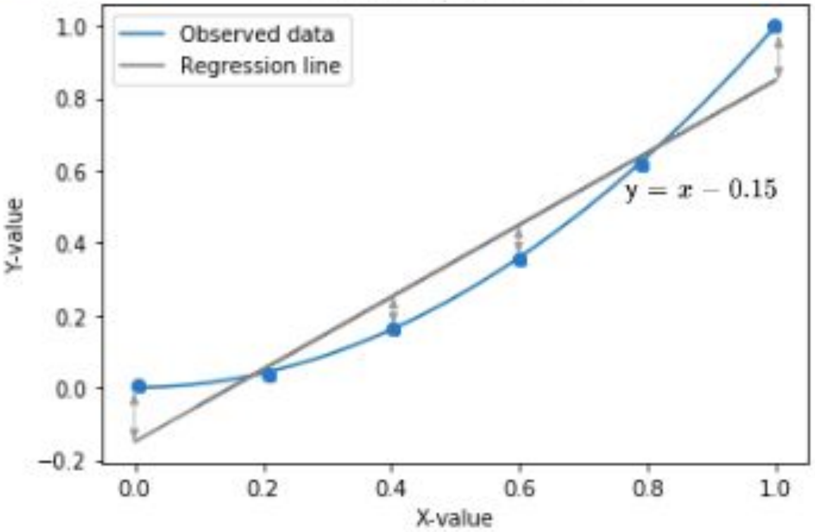
- The sum of squares regression (SSR) is the sum of squared differences between the predicted value and the mean of the response variable

- SSR is the measure of variability in the response variable considering the effect of dependent variable

- It is the explained variation

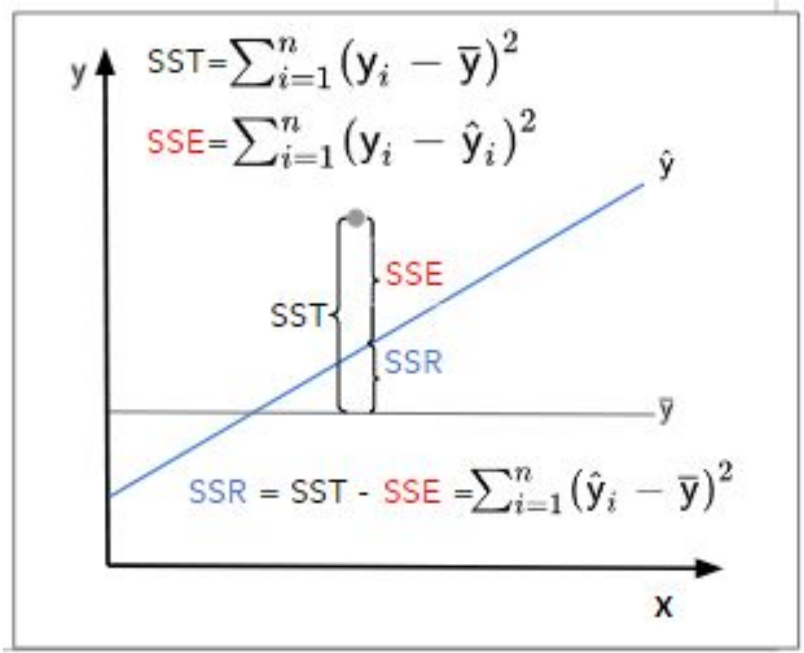
- Also known as Regression Sum of Square (RSS)

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 dependent variable
- It is the unexplained variation
- Also known as Error Sum of Square (ESS)

Variation in response variable


y
$\text{SST}=\sum_{i=1}^{n}(y_i - \bar{y})^2$
$\text{SSE}=\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
ŷ
SSE
SST
SSR
ȳ
$\text{SSR} = \text{SST} - \text{SSE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
x

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 (β_0,β_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.

Inferences about Slope

The t test for significance

- For β to be significant, β > 0.

$$H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta \neq 0$$

- It implies

H_0: The parameter β is not significant

against H_1: The parameter β is significant

- Failing to reject H_0 implies that the parameter β is not significant

The t test for significance

- The test statistic is t given by

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$ where $\hat{\beta}$ is the estimated value of β.

- The t-statistic follows the $t_{(n-2)}$ distribution

- Decision Rule: Reject H_0 if $|t| > t_{(n-2),\alpha/2}$ or if the p-value is less than the α (level of significance)

The t test for slope

- For a existence of a linear relationship $\beta_1 > 0$, to test

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0$$

- It implies

H_0: There is no relationship between variables X and Y

against H_1: There is relationship between variables X and Y

- Failing to reject H_0 implies that there is no relationship between X and Y

The t test for intercept

- For a existence of a linear relationship $\beta_1 > 0$, to test

$$H_0 : \beta_0 = 0 \quad \text{against} \quad H_1 : \beta_0 \neq 0$$

- It implies

H_0: The parameter β_0 is not significant

against H_1: The parameter β_0 is significant

- Failing to reject H_0 implies that the parameter β_0 is not significant

The interval estimation of β

- The interval estimate of a parameter gives the 100(1-α)% confidence interval

(Say α = 0.05, 100(1-α)% = 95%)

- In other words, for an experiment conducted 100 times, the estimate would lie within the confidence interval 95 times. This would give the 95% confidence interval

Interval estimation for slope

- The test statistic for slope is

$$t_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$ where $t_1 \sim t_{(n\text{-}2)}$

- The 100(1-α)% confidence interval for slope is given by

$$(\hat{\beta}_1 - t_{(n-2),\alpha/2} SE(\hat{\beta}_1), \hat{\beta}_1 + t_{(n-2),\alpha/2} SE(\hat{\beta}_1))$$

where $\hat{\beta}_1$ is the estimated value of β_1 and n are the number of observations

Interval estimation for intercept

- The test statistic for slope is

$$t_0 = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}$$ where $t_0 \sim t_{(n-2)}$

- The 100(1-α)% confidence interval for slope is given by

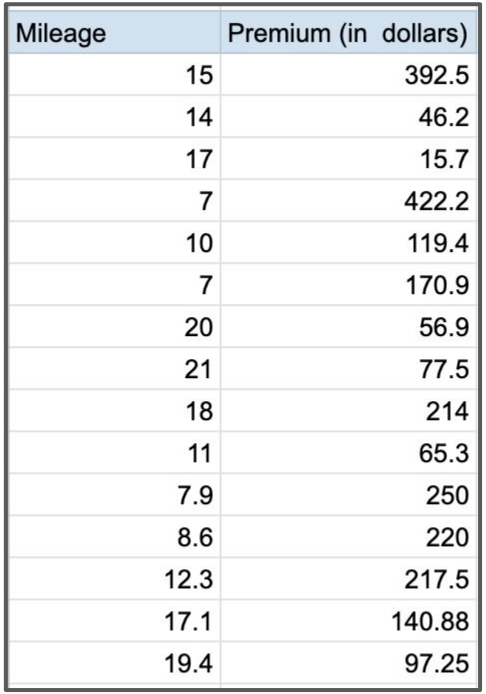
$$(\hat{\beta}_0 - t_{(n-2),\alpha/2} SE(\hat{\beta}_0), \hat{\beta}_0 + t_{(n-2),\alpha/2} SE(\hat{\beta}_0))$$

where $\hat{\beta}_0$ is the estimated value of β_0 and n are the number of observations

Confidence intervals

- We have $\alpha = 0.05$, thus $\alpha/2 = 0.025$
- For the lower bound of CI, $0 + \alpha/2 = 0.025$
- For the upper bound of CI, $1 - \alpha/2 = 0.975$

Parameter	0.025	0.975
β_1	-24.665	1.284
β_0	139.057	515.115



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

ANOVA for regression

- The hypothesis for ANOVA in regression framework are

$H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$

- It implies

H_0: The regression model is not significant

against H_1: The regression model is significant

ANOVA table for bivariate regression

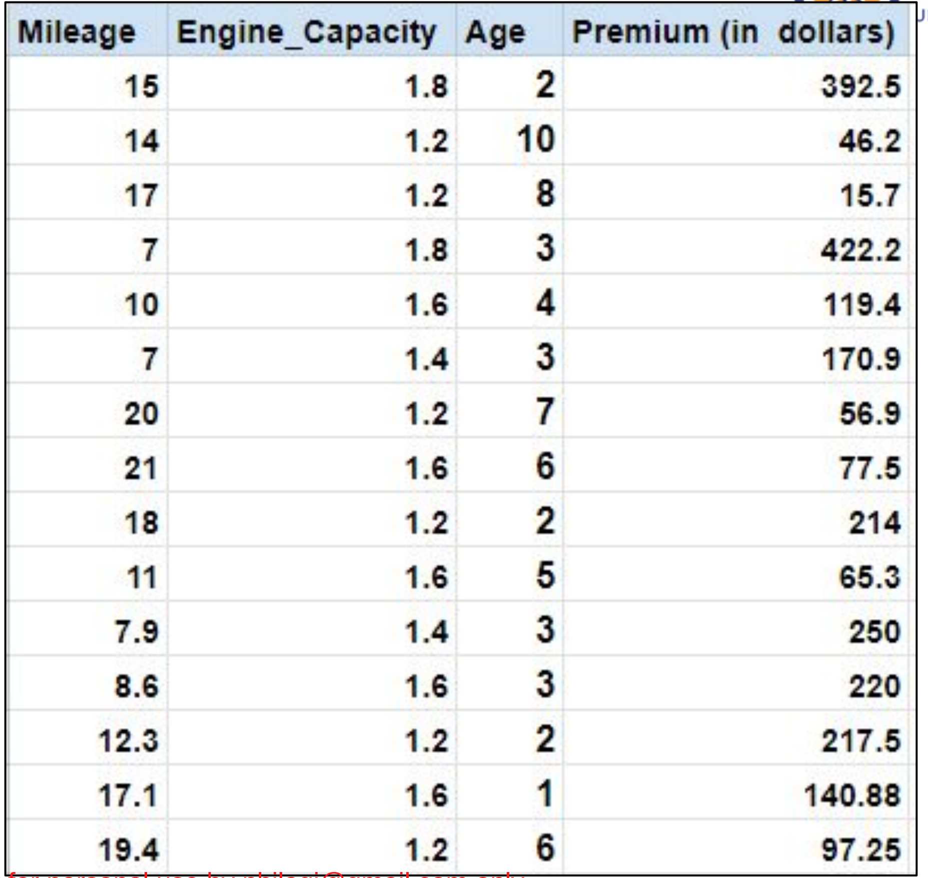
Source of variation	Sum of Squares	Degrees of Freedom	Mean Sum of Squares	F ratio
Regression	RSS	k = 1	MRSS = RSS/1	F_0 = MRSS/MESS
Residual	ESS	n- k - 1 = n - 1 -1 = n -2	MESS = ESS/(n-2)	
Total	TSS	n - 1	-	

- Decision rule: Reject H_0, if $F_0 > F_{(1,n-2),\alpha}$ or if the p-value is less than the α (level of significance)

- Failure to reject H_0 implies that the model is not significant

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

The t test for correlation coefficient

- For a existence of a correlation ρ, i.e. to test

$$H_0 : \rho = 0 \quad \text{against} \quad H_1 : \rho \neq 0$$

- It implies

H_0: There is no correlation

against H_1: The correlation is significant

- Failing to reject H_0 implies that there is correlation

The t test for correlation coefficient

- The test statistic is t_{xy} given by

$$t_{xy} = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

ρ: correlation coefficient
n: number of observations

- The t-statistic follows the $t_{(n-2)}$ distribution

- Decision Rule: Reject H_0 if $|t_{xy}| > t_{(n-2),\alpha/2}$ or the p-value is less than the α (level of significance)

Multiple Linear Regression

Multiple linear regression

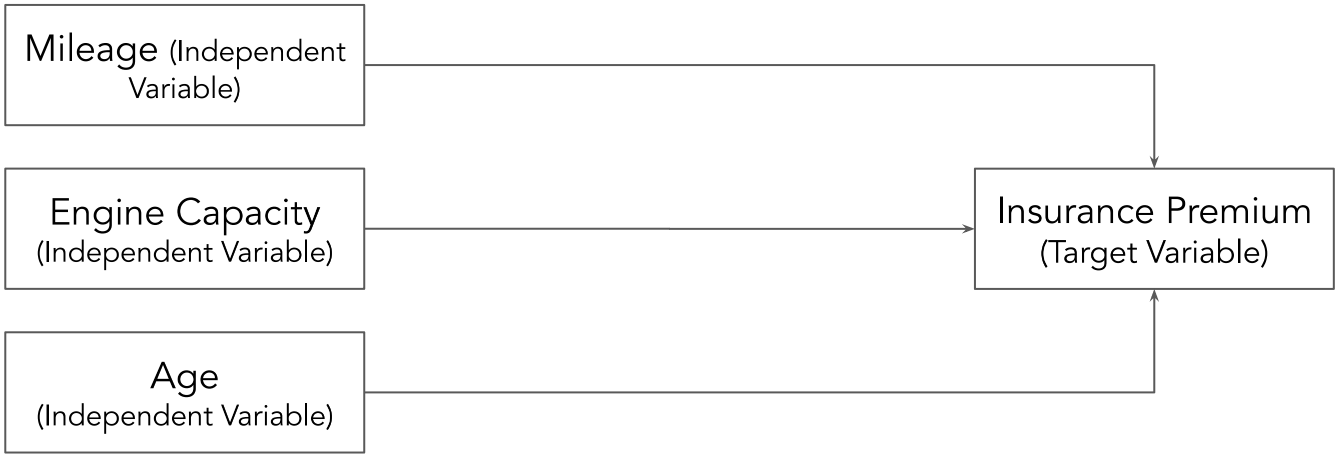
Multiple regression model is used when multiple predictor variables [X_1, X_2, X_3, …, X_n] are used to predict the response variable Y

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

β_0, β_1, β_2, β_3, …, β_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


Mileage (Independent Variable)
Engine Capacity (Independent Variable)
Age (Independent Variable)
Insurance Premium (Target Variable)

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

Linear regression line

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

y = set of values taken by dependent variable Y

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

β_0 = y intercept

β_i = beta coefficient for the ith 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
β_0	Premium value where the best fit line cuts the Y-axis (Premium)
β_1	Regression coefficient of variable Mileage
Mileage	Set of values taken by the variable Mileage
β_2	Regression coefficient of variable Engine_Capacity
Engine_Capacity	Set of values taken by the variable Engine_Capacity
β_3	Regression coefficient of variable Age
Age	Set of values taken by the variable Age
ε	Error component

Parameter estimation - OLS method

- We obtain the estimates of β_0, β_1, β_2 and β_3 to minimize the term

$$E = \sum \boldsymbol{\varepsilon}^{\mathbf{2}} = y - \sum(\mathbf{y} - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)\mathbf{)}^{\mathbf{2}}$$

- To minimize the error we take partial derivatives with respect to β_0, β_1, β_2 and β_3 and equate them to zero

$$\delta E/\delta\beta_0 = 0 \qquad \delta E/\delta\beta_1 = 0$$
$$\delta E/\delta\beta_2 = 0 \qquad \delta E/\delta\beta_3 = 0$$

- So we get four equations with four unknowns, β_0, β_1, β_2 and β_3

Parameter estimation - OLS method

- Solving those equations gets tough
- So, we make use of matrix form, in order to get OLS estimates
- We will first see matrix notation for simple linear regression and then for multiple linear regression

Equations for simple linear regression

Using (x_1, y_1), (x_2, y_2), (x_3, y_3)…… (x_n, y_n) we would have the equations:

$$y_1 = (\beta_0 + \beta_1 x_{11}) + \varepsilon_1$$

$$y_2 = (\beta_0 + \beta_1 x_{12}) + \varepsilon_2$$

$$y_3 = (\beta_0 + \beta_1 x_{13}) + \varepsilon_3$$

$$\ldots$$

$$y_n = (\beta_0 + \beta_1 x_{1n}) + \varepsilon_n$$

Matrix equation for simple linear regression

Expressing the equations from previous slide in matrix form:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{13} \\ \vdots & \vdots \\ 1 & x_{1n} \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

n x 1 n x 2 2 x 1 n x 1

This gives us the Matrix equation: $Y = \beta X + \varepsilon$

Using Linear regression technique, we solve for β's

Equations for multiple linear regression

For 3 predictor variable and n observations, we would have the following equations:

$$y_1 = (\beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \beta_3 x_{31}) + \varepsilon_1$$
$$y_2 = (\beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \beta_3 x_{32}) + \varepsilon_2$$
$$y_3 = (\beta_0 + \beta_1 x_{13} + \beta_2 x_{23} + \beta_3 x_{33}) + \varepsilon_3$$
$$\ldots$$
$$y_n = (\beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \beta_3 x_{3n}) + \varepsilon_n$$

Matrix equation for multiple linear regression

In Matrix form, it would look as follows:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} \\ 1 & x_{12} & x_{22} & x_{32} \\ 1 & x_{13} & x_{23} & x_{33} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

n x 1 n x (3+1) (3+1) x 1 n x 1

Here n is the number of observations.

The OLS estimates

For multiple linear regression, the OLS estimates which give the best fit are obtained as

$$\hat{\beta} = [X'X]^{-1}X'Y$$

X′ denotes the transpose of matrix X.

$$\hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

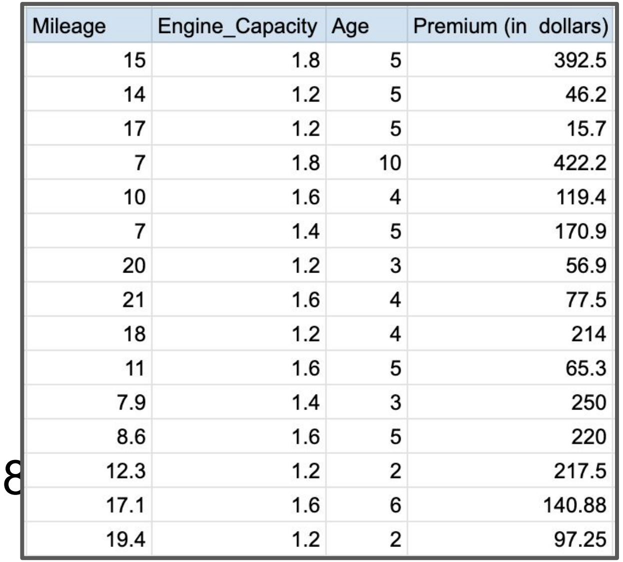
$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} \\ 1 & x_{12} & x_{22} & x_{32} \\ 1 & x_{13} & x_{23} & x_{33} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} \end{bmatrix}$$

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

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

That is,

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

Interpreting the β coefficients

In context with our example,

- β_0 = 138.398: the value of premium when the mileage, engine capacity and age are all equal to 0 (which is absurd)
- β_1 = - 4.876: the average decrease in the premium of the cars due to the mileage
- β_2 = 137.633: the average increase in the premium of the cars due to engine

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

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

ANOVA for regression with ‘k’ predictors

- The hypothesis for ANOVA in regression framework are

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ against H_1: At least one $\beta_k \neq 0$ (k =1,2,3)

- It implies

H_0: the regression model is not significant

against H_1: the regression model is significant

ANOVA table for regression with ‘k’ predictors

Source of variation	Sum of Squares	Degrees of Freedom	Mean Sum of Squares	F ratio
Regression	RSS	k	MRSS = RSS/1	F_0 = MRSS/MESS
Residual	ESS	n - k - 1	MESS = ESS/(n-k-1)	
Total	TSS	n-1	-	

- Decision rule: Reject H_0, if $F_0 > F_{(k,n-k-1),\alpha}$ or if the p-values is less than the α (level of significance)
- Failure to reject H_0 implies that the model is not significant