# Fundamentals of Artificial Intelligence

## Linear Regression

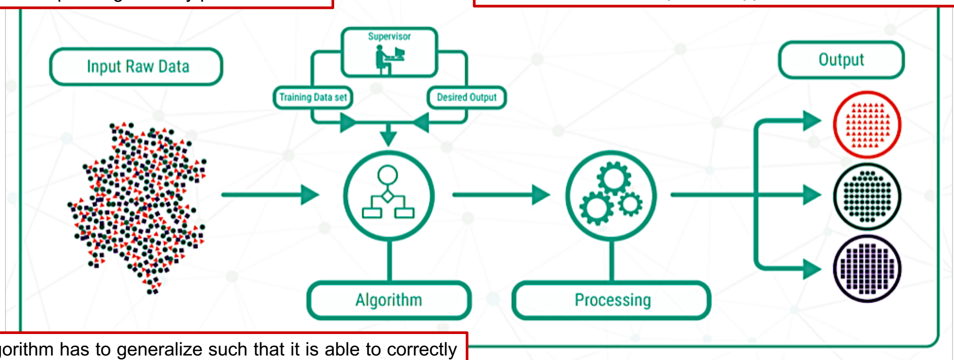
**Shyamanta M Hazarika**
Mechanical Engineering
Indian Institute of Technology Guwahati
s.m.hazarika@iitg.ac.in

http://www.iitg.ac.in/s.m.hazarika/

# Supervised Learning

The set of (training/learning) data consists of a set of input data and correct responses (labels) corresponding to every piece of data.

Supervised learning: a type of machine learning that learns from training data with labels as learning targets. It is the most widely used type of machine learning



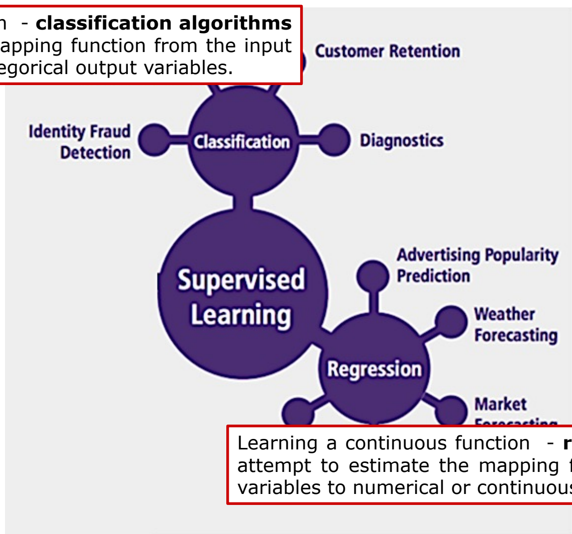

The algorithm has to generalize such that it is able to correctly (or with low error margin) respond to all possible inputs.

Image Source: Data Demystified — Machine Learning; http://towardsdatascience.com

**Learn to predict** output when **given an input vector**

# Supervised Learning

Learning a discrete function - **classification algorithms** attempt to estimate the mapping function from the input variables to discrete or categorical output variables.

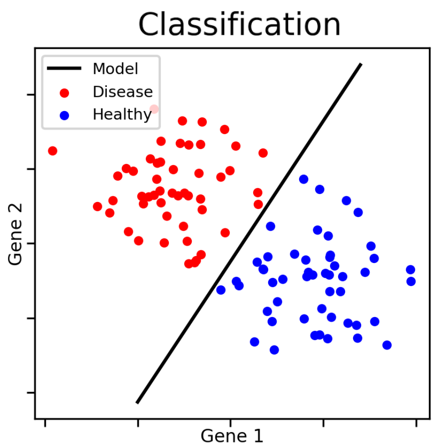



Learning a continuous function - **regression algorithms** attempt to estimate the mapping function from the input variables to numerical or continuous output variables.

Image Source: DHL, Artificial Intelligence in Logistics, 2018.




# Classification vs. Regression




Regression predicts a **continuous target variable Y**. It allows you to estimate a value, such as housing prices or human lifespan, based on input data X.

# Independent and Dependent Variable

- **Independent Variable:**
  A variable whose value does not change by the effect of other variables and is used to manipulate the dependent variable. It is often denoted by X.

- **Dependent Variable:**
  A variable whose value change when there is any manipulation in the values of independent variable. It is often denoted as Y .

Explanatory variables are termed the **independent** variables and the variables to be explained are termed the **dependent** variables.




# Bivariate and Multivariate models

Bivariate or simple regression model

Education X $\longrightarrow$ Y Income

Multivariate or multiple regression model

Education $X_1$

Sex $X_2$ $\longrightarrow$ Y Income

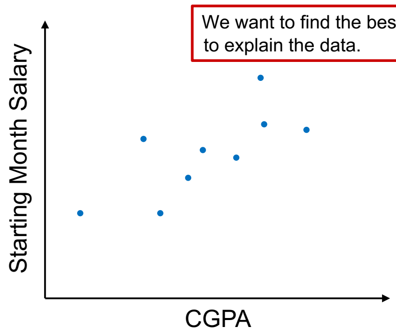
Experience $X_3$

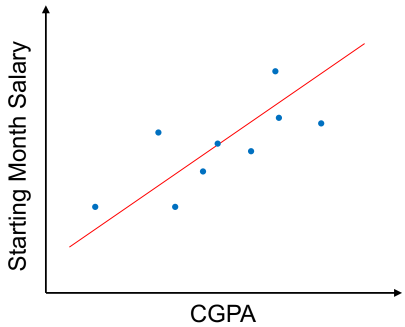
Age $X_4$

# Linear Regression

Relation between variables where changes in some variables may "explain" or possibly "cause" changes in other variables.

We want to find the best line (linear function y=f(X)) to explain the data.




Linear regression is one of the oldest forms of machine learning. It is a long-established statistical technique that involves simply fitting a line to some data.




# Simple Linear Regression Model

- The equation that describes how y is related to x and an error term is called the regression model.
- The simple linear regression model is:

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

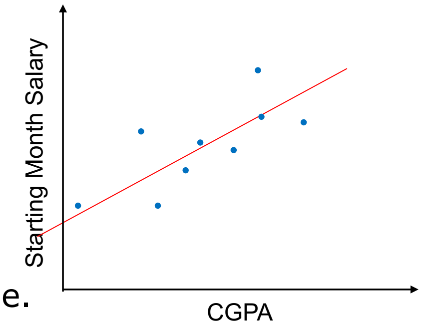
Where $\beta_0$ and $\beta_1$ are the parameters of the model; and ε is the unexplained, random, or error component.

- The two parameters to estimate are the slope of the line and the y-intercept.

# Simple Linear Regression Equation

- The simple linear regression equation is:

$$E(y) = \beta_0 + \beta_1 x$$




1. The regression equation is a straight line.
2. $\beta_0$ is the y intercept of the regression line.
3. $\beta_1$ is the slope of the regression line.
4. $E(y)$ is the expected value of y for a given value of x.




# Estimated Linear Regression Equation

- The estimated simple linear regression equation is

$$\hat{y} = b_0 + b_1 x$$

- $\hat{y}$ refers to the predicted values of the dependent variable y that are associated with values of x, given the linear model.
- From the sample of values of x and y, $b_0$ estimate of $\beta_0$ and $b_1$ estimate of $\beta_1$ are obtained.

# Estimation Process

Regression Model
$y = \beta_0 + \beta_1 x + \varepsilon$
Regression Equation
$E(y) = \beta_0 + \beta_1 x$
Unknown Parameters
$\beta_0, \beta_1$



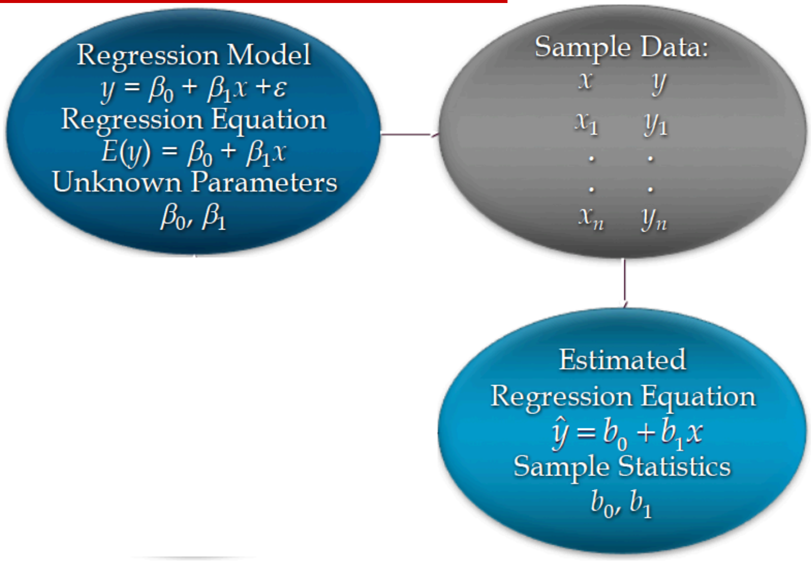

# Estimation Process

Regression Model
$y = \beta_0 + \beta_1 x + \varepsilon$
Regression Equation
$E(y) = \beta_0 + \beta_1 x$
Unknown Parameters
$\beta_0, \beta_1$

Sample Data:

| $x$ | $y$ |
|---|---|
| $x_1$ | $y_1$ |
| . | . |
| . | . |
| $x_n$ | $y_n$ |

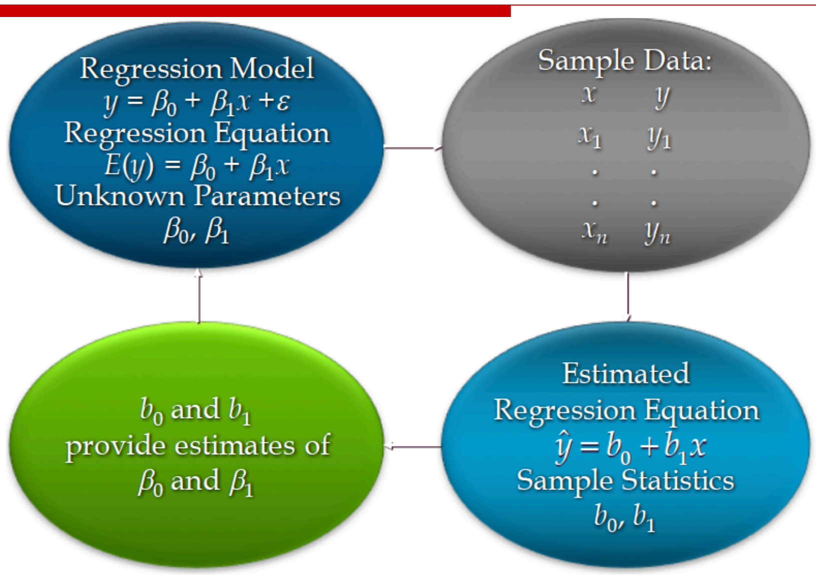
# Estimation Process




13



# Estimation Process




14

# Estimation Process

- Our goal is to learn the **model parameters** that minimize error in the model's predictions.
- To find the best parameters:
  1. Define a **cost function**, or **loss function**, that measures how inaccurate our model's predictions are.
  2. Find the parameters that **minimize loss**, i.e. make our model as accurate as possible.

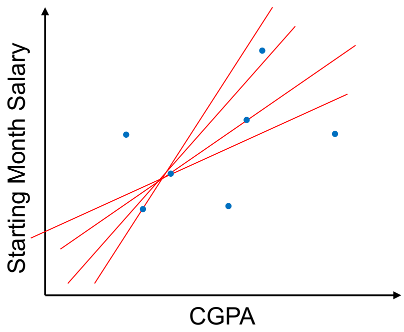
**Gradient descent: learn the parameters**
The goal of gradient descent is to find the minimum of our model's loss function by iteratively getting a better and better approximation of it.

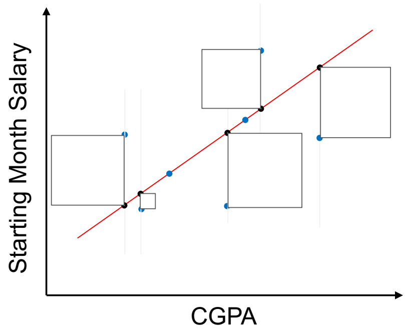



# Estimation Process

- A large number of procedures have been developed for parameter estimation and inference in linear regression.
- These methods differ in computational simplicity of algorithms, presence of a closed-form solution, robustness with respect to heavy-tailed distributions, and theoretical assumptions needed to validate desirable statistical properties such as consistency and asymptotic efficiency.
- One of the most common estimation techniques for linear regression is Least Square Estimation.

# Estimation Process



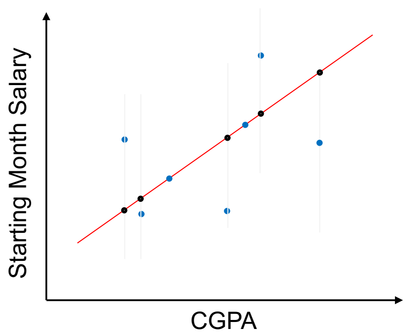

The least-squares regression line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

17



# Least Square Regression

The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve. Least squares regression is used to predict the behavior of dependent variables.

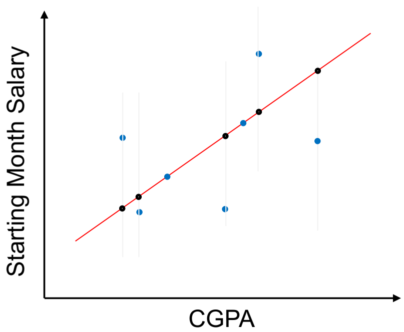



The least-squares regression line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

18

# Least Square Regression

The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve. Least squares regression is used to predict the behavior of dependent variables.




## Least Squares Criterion

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

where:

$y_i$ = observed value of the dependent variable for the i$^{th}$ observation.

$\hat{y}_i$ = estimated value of the dependent variable for the i$^{th}$ observation.





# Least Square Regression

## Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

The **slope** of the regression line describes how much we expect $y$ to change, on average, for every unit change in $x$.

where:

$x_i$ = value of independent variable for i$^{th}$ observation.

$y_i$ = value of dependent variable for i$^{th}$ observation.

$\bar{x}$ = mean value of independent variable.

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

## $y$-Intercept

$$b_1 = (\bar{y} - b_1 \bar{x})$$

The **intercept** is a necessary mathematical descriptor of the regression line. It does not describe a specific property of the data.

# Coefficient of Determination

- r², the coefficient of determination, is the square of the correlation coefficient.
- r² represents the fraction of the variance in y that can be explained by the regression model.
- Mathematically, it can be shown that r² is equal to the fraction of the total sum of squares that is due to the regression model.

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{Sum of squares due to regression}}{\text{Total sum of sqaures}}$$

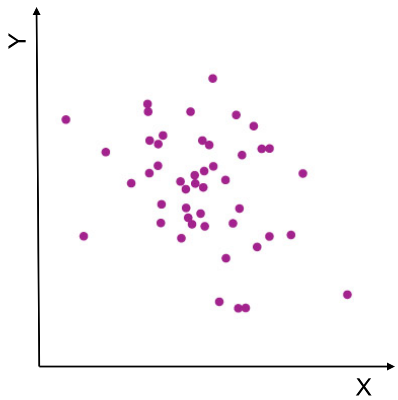
Described as the fraction of the total variance not explained by the model.




# Correlation

- r² represents the fraction of the variance in y that can be explained by the regression model.
- r, the correlation quantifies the strength and direction of a linear relationship between two quantitative variables.
- r is positive for positive linear relationships, and negative for negative linear relationships.
- The closer r is to zero, the weaker the linear relationship is; beware that r has this particular meaning for **linear** relationships only.

# Example - Correlation




*r* = –0.3,

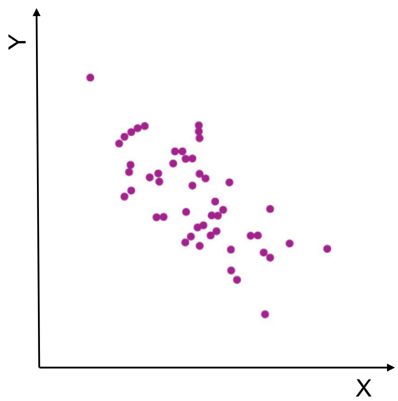
$r^2$ = 0.09, or 9%

The regression model explains not even 10% of the variations in *y*.





# Example - Correlation

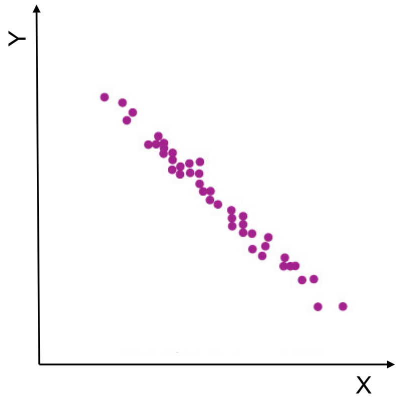



*r* = –0.7,

$r^2$ = 0.49, or 49%

The regression model explains nearly half of the variations in *y*.

# Example - Correlation




$r = -0.99$,

$r^2 = 0.9801$, or 98%

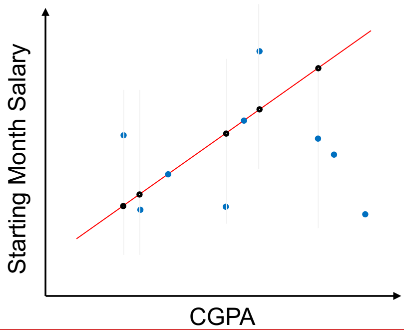
The regression model explains almost all of the variations in $y$.




# Outliers and Influential points

**Outlier:** An observation that lies outside the overall pattern.

**Influential individual:** Observation that markedly changes the regression if removed; often an isolated point.



# Before Making Predictions

- Linear regression is been studied at great length, and there is a lot of literature on how the data must be structured to make best use of the model.
- Rules of thumb for preparation of data when using Ordinary Least Squares Regression, the most common implementation of linear regression.

1. **Linear Assumption**. Linear regression assumes that the relationship between the input and output is linear. It does not support anything else.
   a. This may be obvious, but it is good to remember!
   b. May need to transform data to make the relationship linear (e.g. log transform for an exponential relationship).




# Before Making Predictions

2. **Remove Noise**. Linear regression assumes that the input and output variables are not noisy.
   a. Consider using data cleaning operations that let you better expose and clarify the signal in your data.
   b. This is most important for the output variable and you want to remove outliers in the output variable (y) if possible.

3. **Remove Collinearity**. Linear regression will over-fit the data when you have highly correlated input variables.
   a. Consider calculating pairwise correlations for your input data and removing the most correlated.

# Before Making Predictions

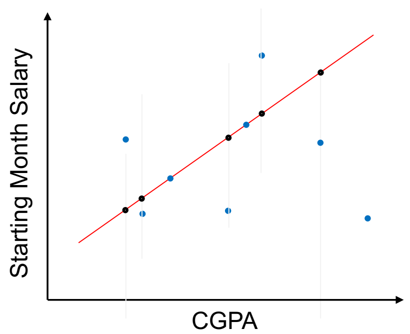
4. **Gaussian Distributions**. Linear regression will make more reliable predictions if the input and output variables have a Gaussian distribution.
   a. May get some benefit using transforms (e.g. log) on the variables to make their distribution more Gaussian looking.

5. **Rescale Inputs**: Linear regression will often make more reliable predictions if you rescale input variables using standardization or normalization.
   a. Feature Scaling or Standardization: It is a step of Data Pre Processing which is applied to features of data. It basically helps to normalize the data within a particular range.




# Making Predictions

- Use the equation of the least-squares regression to predict y for any value of x within the range studied.
- Predication outside the range is extrapolation. Avoid extrapolation.



# Polynomial Regression

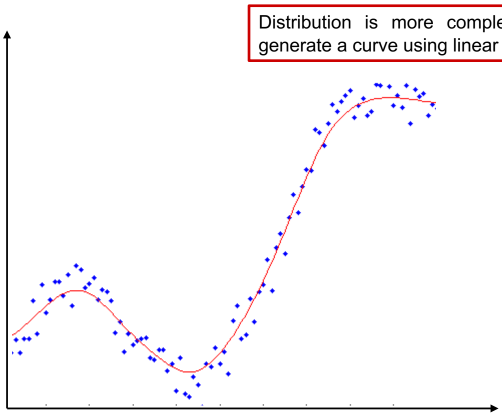
- Polynomial regression is a special case of multiple linear regression analysis.
  - The relationship between the independent variable x and the dependent variable y is modelled as an $n^{th}$ degree polynomial in x.
- In other words, when our data distribution is more complex than a linear one, and we generate a curve using linear models to fit non-linear data.

$$y = W_1x^3 + W_2x^2 + W_3x + W_4$$




# Polynomial Regression

Distribution is more complex than a linear one, and we generate a curve using linear models to fit non-linear data



The independent (or explanatory) variables resulting from the polynomial expansion of the predictor variables are known as higher-degree terms. It has been used to describe nonlinear phenomena such as the growth rate of tissues and the progression of disease epidemics.

# Multivariate Regression

- When each piece of training data is a vector of several attributes, the problem becomes more complicated.
- What we're going to do is to expand the single-attribute example by expanding the dimensions of the space.

For example:

If each example has just two attributes, we could view each labeled example in three-dimensional space, with an x-coordinate corresponding to the first attribute, the y-coordinate corresponding to the second attribute, and the z-coordinate corresponding to the label. We'd look for a plane that minimizes the sum-of-squares error.




# Multivariate Regression

- For a general number of attributes d, we view each labeled example as a point in (d+1)-dimensional space, with a coordinate for each attribute, plus a coordinate for the label.
- We look for a d-dimensional hyperplane that minimizes the sum-of-squares error.
- This is slightly different than the two-dimensional case because the hyperplane is forced to go through the origin. Forcing this keeps the mathematics prettier.

# Final Comments

- Linear regression is really best suited for problems where the attributes and labels are all numeric and there is reason to expect that a linear function will approximate the problem.

- This is rarely a reasonable expectation— linear functions are just too restricted to represent a wide variety of hypotheses.

Oversimplifies the classification rule. (Why should we expect a hyperplane to be a good approximation?)

- Linear regression is widely used in biological, behavioral and social sciences to describe possible relationships between variables.