# Introduction to Soft Computing and Intelligent Systems

# Outline

- **Course Admin**
  - Announcements
  - Evaluation
- **Intelligent Systems**
- **Example: Linear Regression models**

# Introduction to Soft Computing and Intelligent Systems

- Instructor: Haitham Amar- hamar@uwaterloo.ca
- Material: learn.uwaterloo.ca
- Lectures:In-Person format.
  - Tuesday from 5:30 pm- 8:20 pm, E7 4043
  - Thursday from 5:30 pm- 8:20 pm, E7 4043
- TA: Fatemeh Tavakoli- fatemeh.tavakoli@uwaterloo.ca
- TA Office Hours: email Fatemeh to arrange time for that week

# Announcements

- Registering/Waiting List- No Auditing of this course
- Log on to learn.uwaterloo.ca
  - enable email notifications
  - use the message boards, let me know if you want specific groups or categories, I can create them
  - talk to each other

# Work load and Evaluation- Subject to change

- Assignments 40%
  - Assignments will be provided on a biweekly to tri-weekly basis. Students are highly encouraged to work and solve them
  - The assignments are going to require group effort. Please join a group ASAP.
- Exams 60%
  - First Exam 60%: Format (In-Person)

# Required Background

- ECE 650 or equivalent is strongly recommended.
- Math and Linear Algebra : sets, matrices, transpose, cross product, dot product, matrix multiplication, solving system of linear equations
- Programming :
  - You should be comfortable programming in some language, not large software applications but lots of calculations, plotting, etc.
- Probability and Statistics : (not required, we will define or review these, but it would help)
  - definition of probability, Subjective probability, information, entropy, . . .

# Computing Resources

- Course Website:
  - No personal website- ¯_(ツ)_/¯
  - See **Computing Resources** page on website with tips on servers/systems you can use on campus.
- Sharcnet/Compute Canada
  - research students could have supervisor sponsor them to use Sharcnet, no cost.
- If you find useful resources, add to them the resources discussion forum on LEARN.

# Other Tools and Resources

- Mendeley.com Community - online resource for academic papers. Course Group join and post your own papers or comments.
- Kaggle Competition - https://www.kaggle.com/datasets
- Cloud Services - free to use for single user, single machine smaller runs.
  - These have everything we'll cover in this course, we'll learn how to use them, why they are used, to allow you to go beyond them
  - Amazon Web Services (AWS)
  - Azure Tools - Microsoft

# Tools for Data Management and Analysis

- **Only** one tool for this course
- python (The de facto programming choice for data scientists)
  - numpy, scipy, scikit-learn
  - Lots of resources online, communities, modules, new code tools all the time.

# Course Scope and Structure

- The course is useful for graduate students in virtually all areas of engineering, particularly for those dealing with complex systems or processes.
- A background in two or more of the following areas should be useful: fuzzy logic, artificial neural networks, machine learning, AI, system's optimization, nonlinear mapping, calculus of variation, differential calculus, statistical analysis, advanced algebra, game theory.

# Artificial Intelligence

**Definition:**

Algorithms enabled by constraints, exposed by representations that support models, and targeted at thinking, perception, and action.

Without loss of generality, an intelligent system is one that generate hypotheses and test them.

# Intelligent Systems: What Are They?

**Definition:**

Intelligent systems are artificial entities involving a mix of software and hardware which have a capacity to acquire and apply knowledge in an ”intelligent” manner and have the capabilities of perception, reasoning, learning, and making inferences (or, decisions) from incomplete information.

# Intelligent Systems: History

- Al-Jazari's Automata: $12^{th}$ century– He described 100 mechanical devices one of which was a humanoid automata.
- Formal Reasoning: *Circa*$13^{th}$ century
- First mechanical computer: $19^{th}$ century– Charles Babbage
- First program: Ada Lovelace– She wrote, in the $19^{th}$ century, set of notes that completely detail a method for calculating Bernoulli numbers

# Intelligent Systems: History (Cont.)

- The Turing Test: 1950.
- Machine Learning algorithms: started at the mid $20^{th}$
- James Robert Slagle: A symbolic integration program
- Mycine: a backward chaining expert system that used artificial intelligence to identify bacteria causing severe infections– Circa 1970
- Deep Blue of IBM
- Deep Learning: Lots of cool and practical applications,

For example,
https://player.vimeo.com/video/192179726?color=cc0000title=0byline=
192179726

# Intelligent Systems: Features

## Features

A feature that is indispensable in these systems is the generation of outputs, based on some inputs and the nature of the system itself. The inputs to a system may include information as well as tangible items, and the outputs may include decisions as well as physical products.

# Intelligent Systems: Capabilities

It is commonly accepted that an intelligent system possesses one or more of the following characteristics and capabilities:

## Capabilities

Sensory perception; Pattern recognition; Learning and knowledge acquisition; Inference from incomplete information; Inference from qualitative or approximate information; Ability to deal with unfamiliar situations; Adaptability to new, yet related situations (through expectational knowledge); Inductive reasoning.

# Example

A typical input variable is identified for each of the following examples of dynamic systems:

- Human body: neuroelectric pulses
- Company: information
- Power plant: fuel rate
- Automobile: steering wheel movement
- Robot: voltage to joint motor

# Example (Cont.)

Possible output variables for each of these systems are:

- Human body: Muscle contraction, body movements
- Company: Decisions, finished products
- Power plant: Electric power, pollution rate
- Automobile: Front wheel turn, direction of heading
- Robot: Joint motions, effector motion

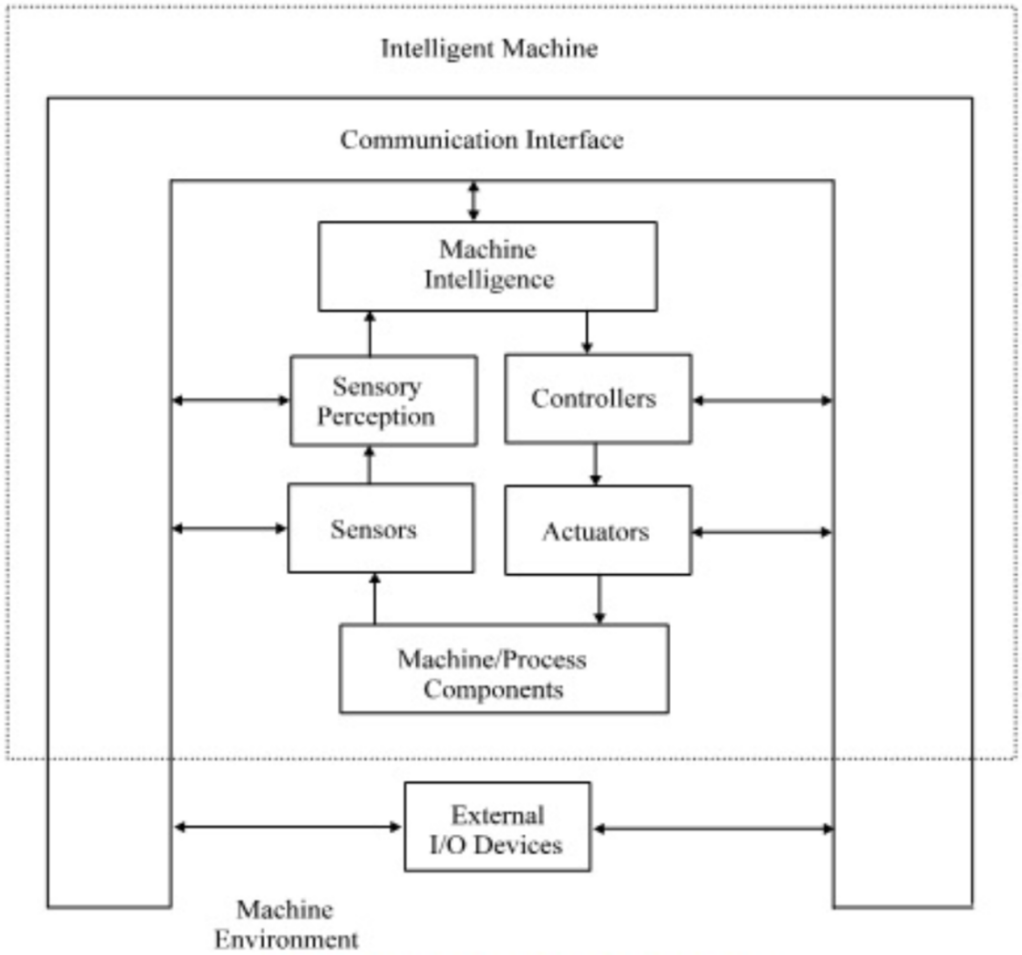
# Intelligent Machines

## Definition

An intelligent machine is a machine that can exhibit one or more intelligent characteristics of a human. As much as neurons themselves in a brain are not intelligent but certain behaviours that are affected by those neurons are, the physical elements of a machine are not intelligent but the machine can be programmed to behave in an intelligent manner.

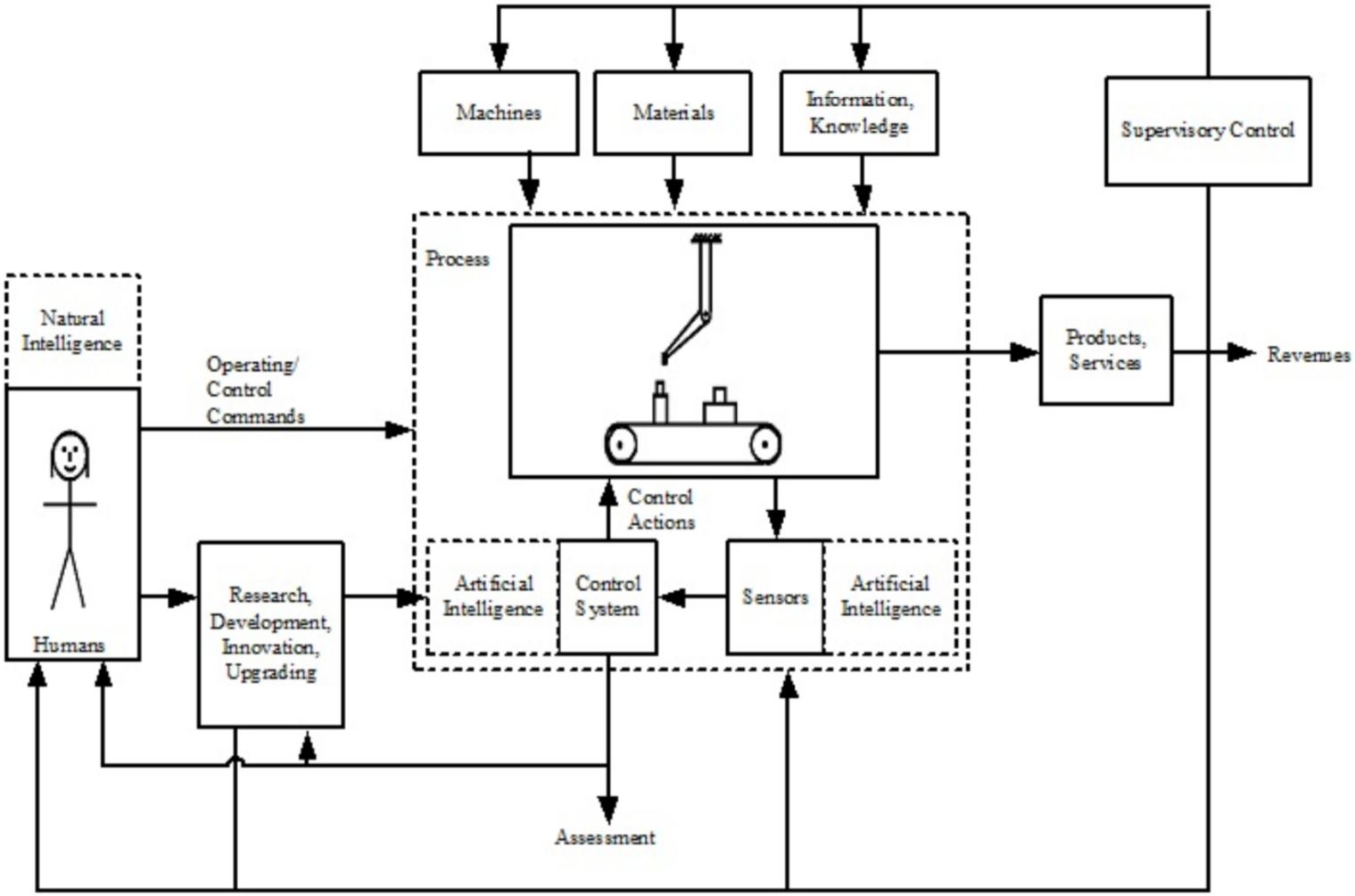
An intelligent machine embodies machine intelligence. An intelligent machine, however, may take a broader meaning than an intelligent computer.

# Intelligent Machines: Architecture




An intelligent Machine

# Intelligent Machines: Schematics and Modules



# What is an intelligent System

If a system solves the following problem, is it intelligent?

$$\int \frac{-5x^4}{(1-x^2)^{\frac{5}{2}}} dx$$

Let's see if this is true?

# What is an intelligent System

Safe transformations:

- $\int cf(x)dx = c\int f(x)dx$
- $\int f(x)dx = \int f(x)dx$
- $\int -f(x)dx = -\int f(x)dx$
- $\int \frac{P(x)}{Q(x)}dx \rightarrow$ Divide

# What is an intelligent System

Heuristic transformations:

- A) f(sin x, cos x, tan x, cot x, sec x, cosec x)
  - g(sin x, cos x)
  - g(tan x, cosec x)
  - g(cot x , sec x)
- B) $\int f(\tan x)dx = \int \frac{f(y)}{1+y}dy$
- c) Use roles such as $\sin^2 x + \cos^2 x = 1$
  - $1 - x^2 \rightarrow x = \sin y$
  - $1 + x^2 \rightarrow x = \tan y$

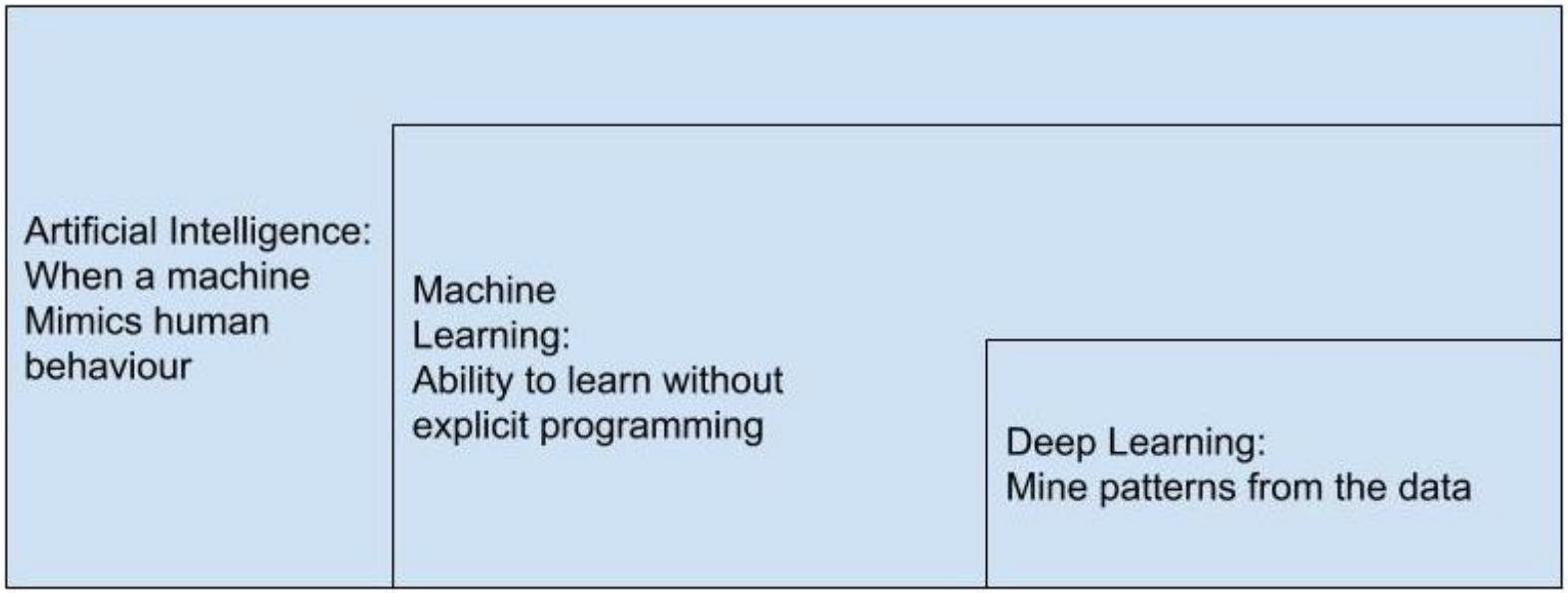
# What is an intelligent System

To solve this problem, you need to :

- Define your goal(s).
- Define the possible transformations.
  - Safe transformations.
  - heuristic transformations.
- Use goal trees to proceed with the solution.
  - What is a Goal Tree?
- When in doubt, use functional composition depth as a way of choosing the best path.
  - What is a functional composition depth?

# Intelligent Systems and Machine Learning

Intelligent systems use artificial intelligence and machine learning. This helps machines to “learn” in much the same way humans do. This is now possible because of ubiquitous data in the modern world, including the ability to store it and communicate it at high speeds.



# Inspirations

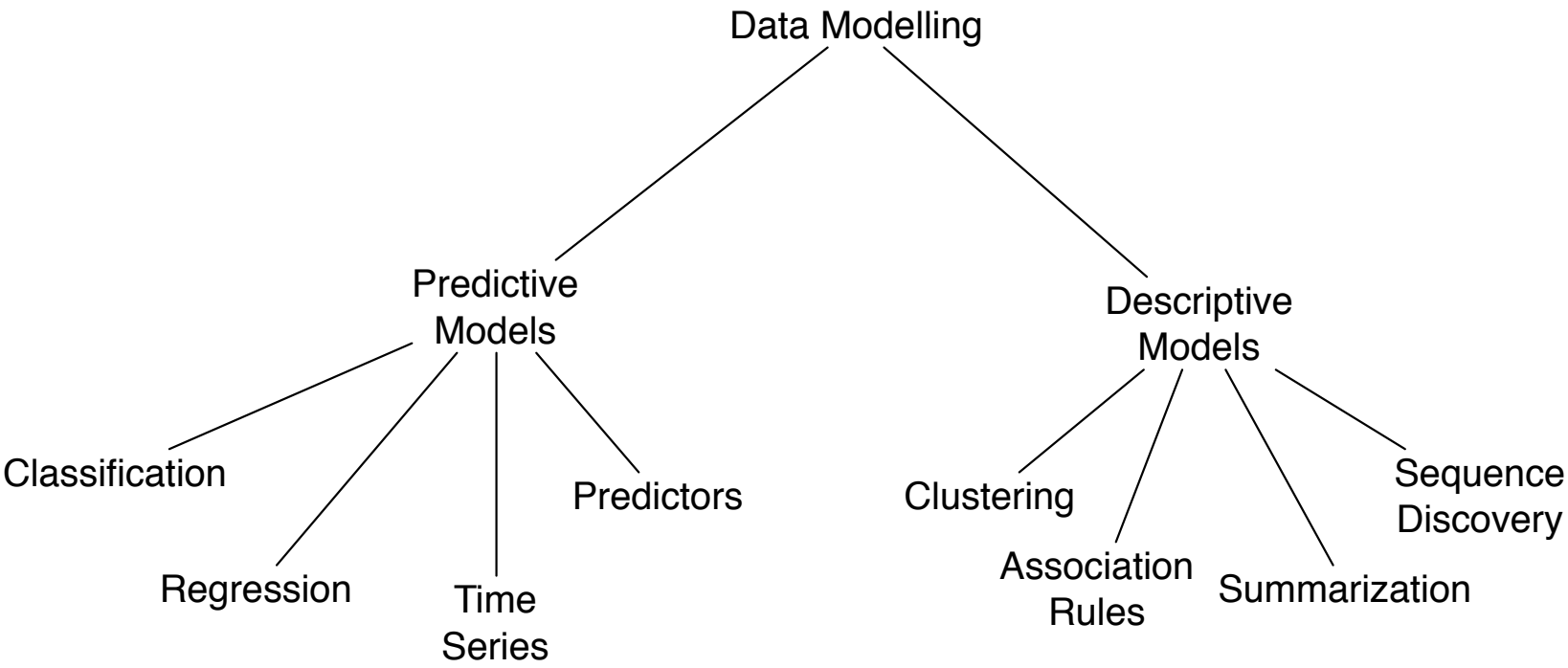
- Models of Problem Solving
- Models of Learning
- Neural Nets
- Experts Based systems
- Evolutionary-like systems

# What do we want to do?

- Have machines behave like humans
- Perform intelligence sequences of decision making
- Do in a fast and efficient way
- etc.

# Where to Start?

- Work with Machine learning algorithms



# Descriptive versus Inferential Analysis

- We have data (samples). This data is a sample of a population (more than just the measured or observed sample).
- **Inferential Analysis (predictive)** is the type of analysis that can describe measures over the population of data. That is observed and unobserved. Fuzzy inference systems use human reasoning to allow for the prediction to happen.
- **Descriptive Analysis** is the type of analysis and measures that seek to describe and summarize the data, the available samples. We can not in general use it for interpretation of unobserved data.

# Major Categories of Predictive modelling

- Regression: Map inputs to an output $\in \mathbb{R}$
- Classification: Map inputs to a categorical output

# Simple Linear Regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted $X$, is regarded as the **predictor**
- The other variable, denoted $Y$, is regarded as the **response**

Different types of relationships between two variables (you can generalize these definitions to any arbitrary set variables):

- deterministic (or functional) relationships: when there is a mathematical function that defines y based on x. For example, the relation between Fahrenheit and Celsius degrees

$$\text{Cels} = \frac{5}{9}\text{Fahr} - 32 \tag{1}$$

# Simple Linear Regression

- Statistical relationships : when there is not mathematical formula (function) that relates y to x.
- Example for statistical relation: Consider a data set in which the Mort variable is the mortality due to skin cancer (number of deaths per 10 million people) and the lat variable is the latitude of the centre of each state. We consider Mort as y, response, and lat as x, predictor.

# Simple Linear Regression

- In the case where we are dealing with statistical relation, we need to use statistical methods to estimate a mathematical relation that best fit the data set (can describe the data set in an efficient way)
- Let $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$ be the sample that we observe by conducting an experiment. We define the covariance of Y and x to be

$$\text{Cov} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \tag{2}$$

# Simple Linear Regression

Next we use the Correlation definition as follows

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{S_x S_y} \tag{3}$$

where $S_x$ and $S_y$ are the estimates of standard deviations for the $X$ observations and $Y$ observations, respectively.

# Simple Linear Regression

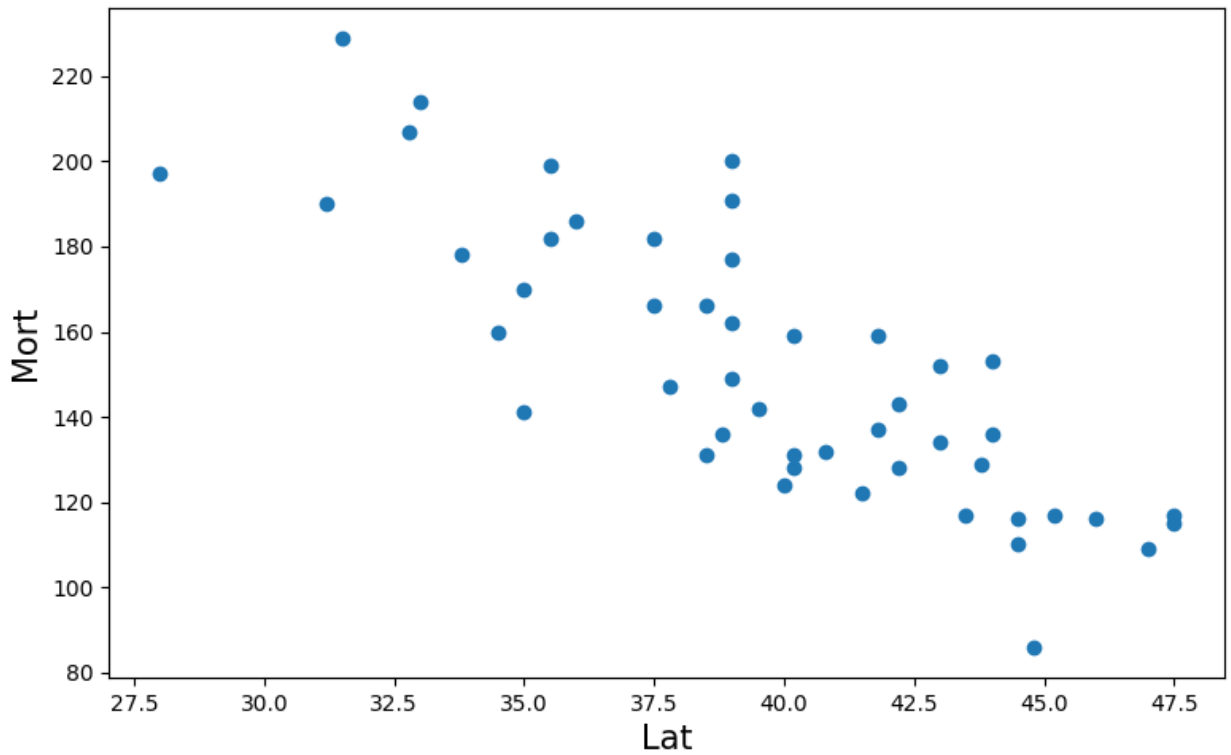
- Fact:

$$-1 \leq \text{Cor}(X, Y) \leq 1 \tag{4}$$

- In general, we have two types of statistical relations:
  1. Linear relation: If Cor(Y,X) is close to 1, it means that we have a positive linear relation (The slop is positive). If Cor(Y,X) is close to -1, it means that we have a negative linear relation.
  2. Non-linear relation: If Cor(Y,X) is close to zero, it means that we do not have any linear relation.

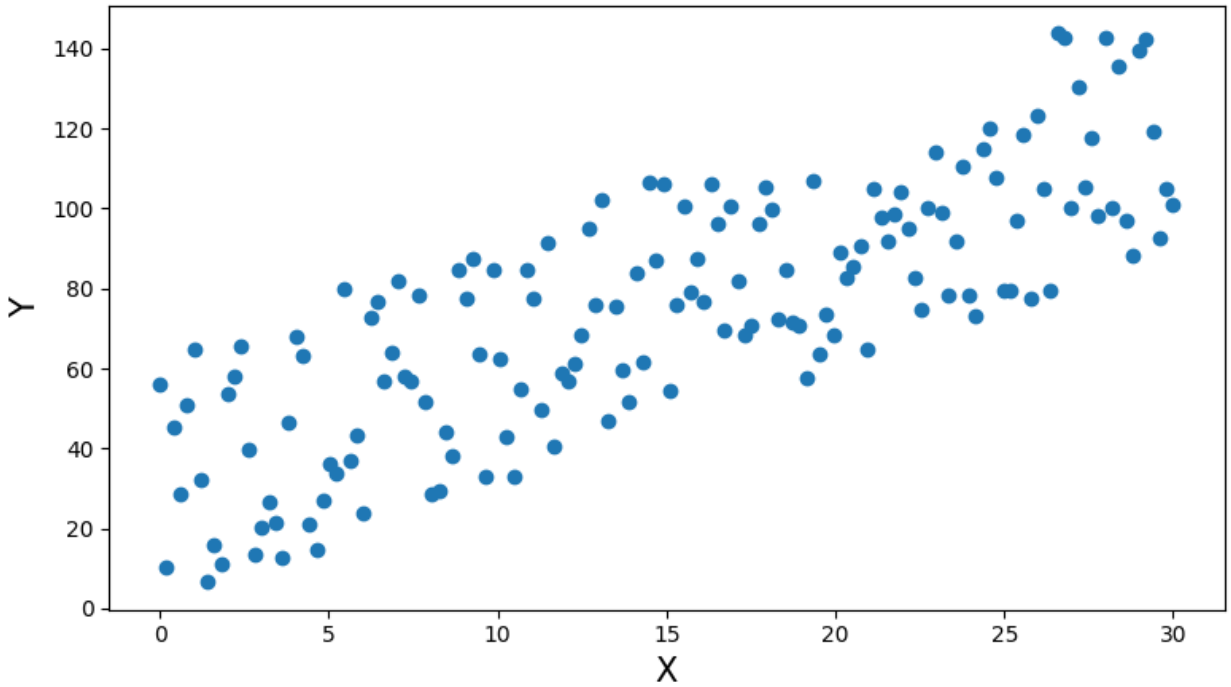
# Simple Linear Regression

Example of Linear Relation (decreasing): cor(y,x)=-0.81



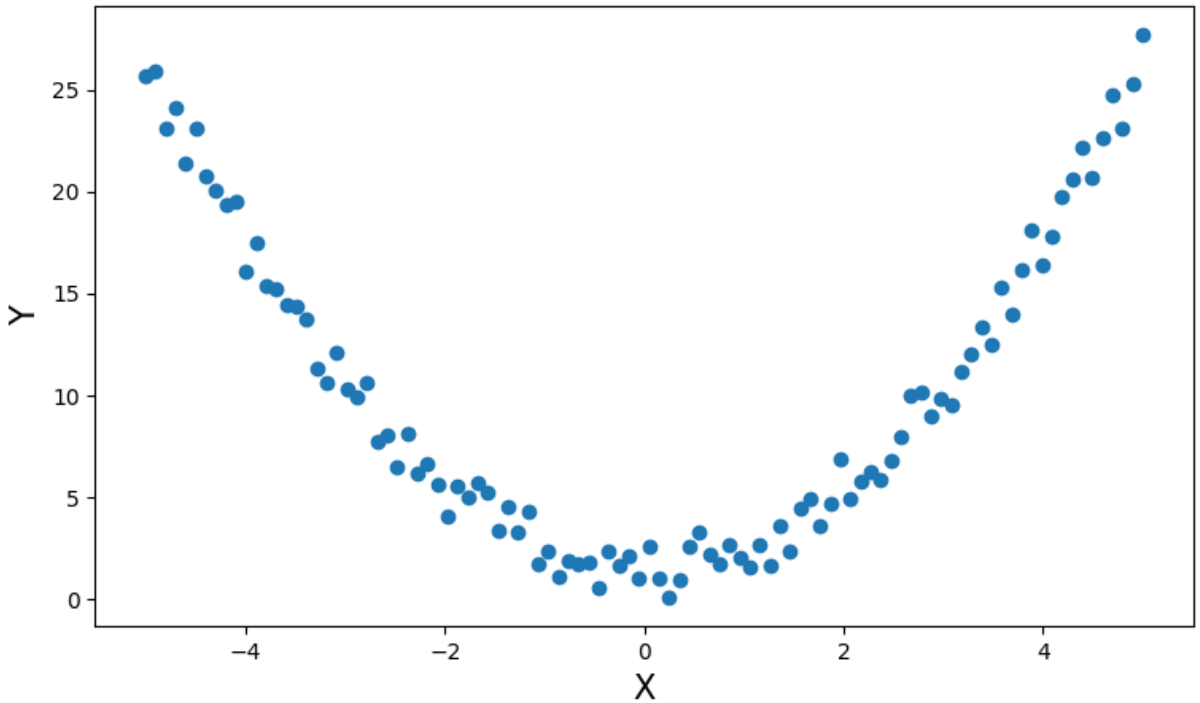
# Simple Linear Regression

Example of Linear Relation (increasing): cor(y,x)=0.824



# Simple Linear Regression

Example of None-Linear Relation: cor(y,x)=0.002

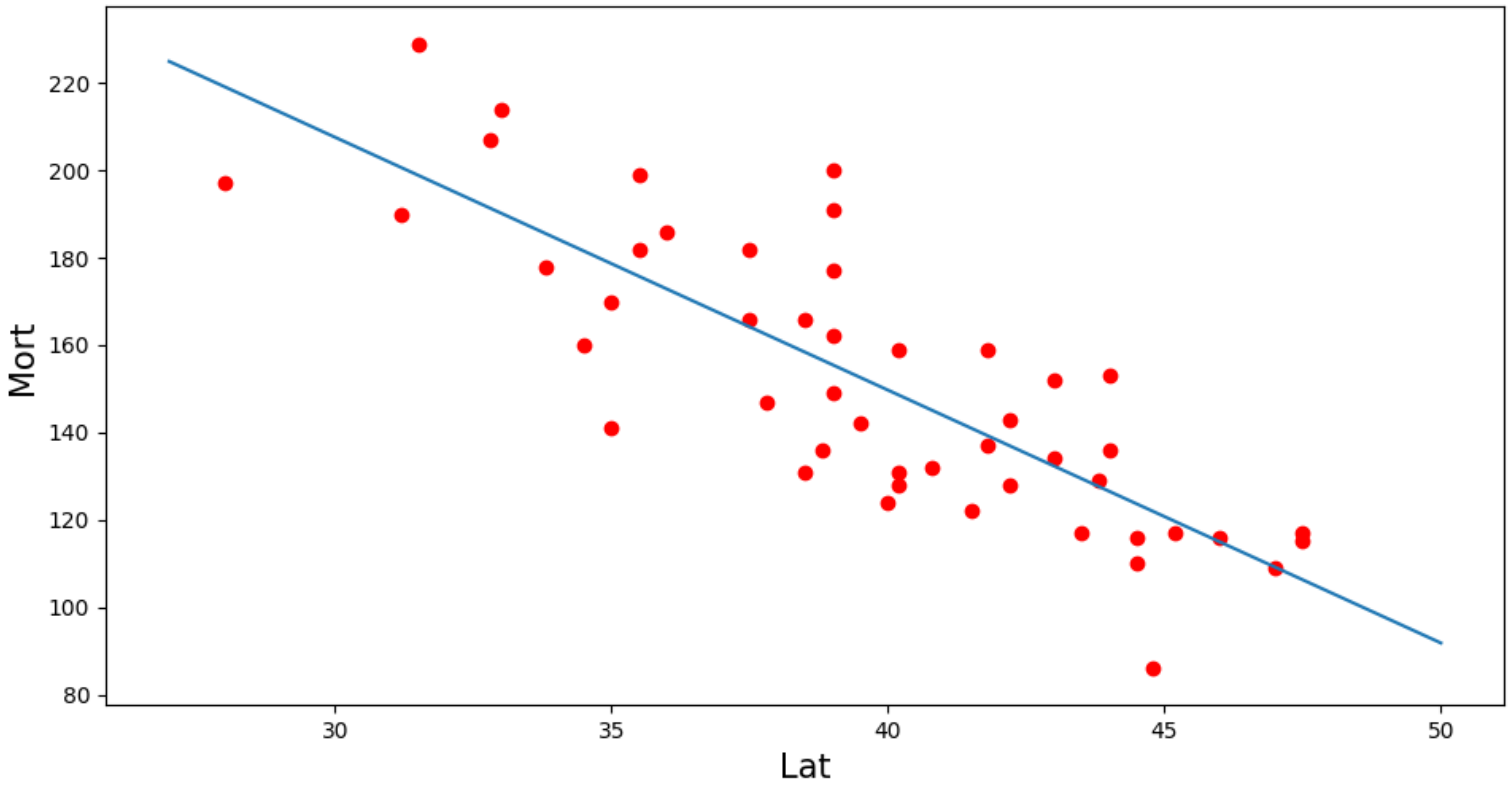


# Simple Linear Regression

- Suppose we conduct an experiment and in this experiment we have two variables, y and x. Suppose that y depends on x and y has a linear relation with x (cor(x,y) is close to 1 or -1). Since y has a linear relation with x, we can find an equation of a line that can describe our data set.
- The line that can describe our data set in the most efficient way is called linear regression line or model. If the line is describing the relation between two variables (y and x), it is called simple linear regression model or line.

# Simple Linear Regression

Best Fit:



# Simple Linear Regression

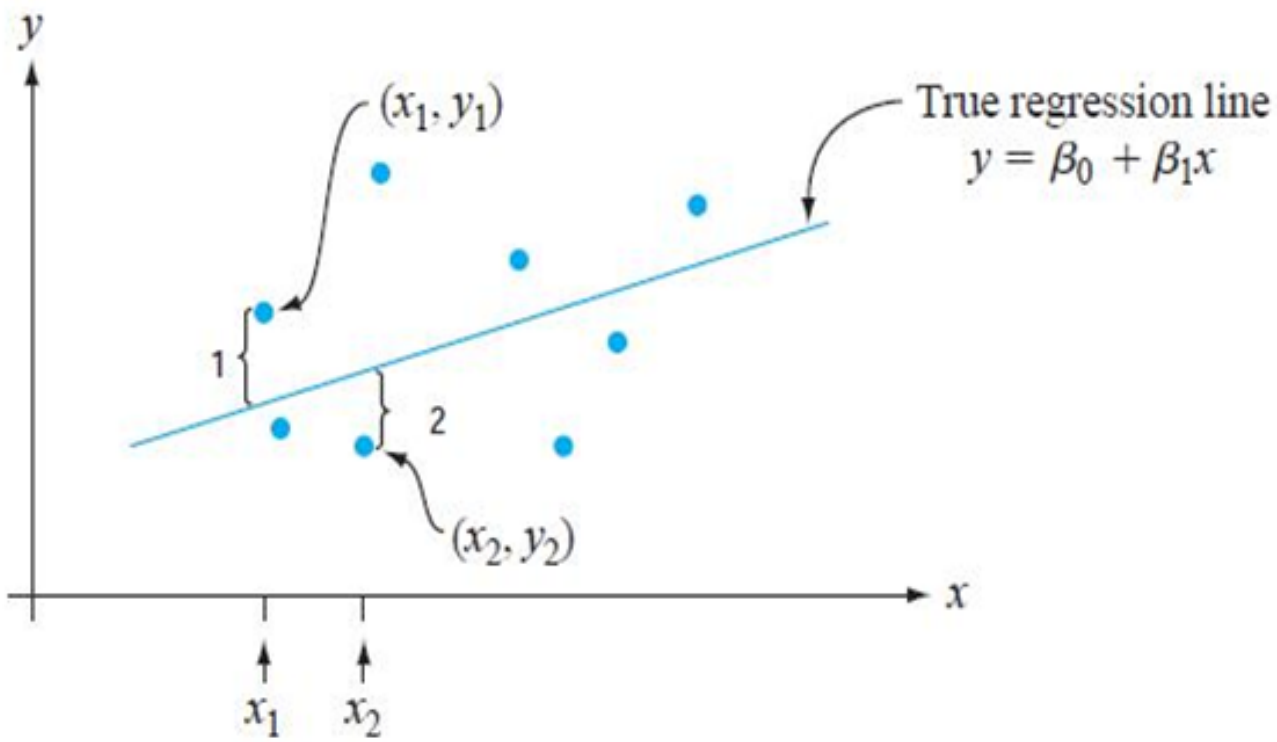
How to find the best fitted line?

- Suppose
  1. $y_i$=denotes the observed response for experimental unit $i$
  2. $x_i$=denotes the predictor value for experimental unit $i$
- We are looking for a line $\hat{y}_i = \beta_0 + \beta_1 x_i$, such that

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

is minimum where $\hat{y}_i$ denotes the predicted response

# Simple Linear Regression



# Theory of Linear Regression

- Maximum Likelihood and Model Estimation
- Assume the probability of observing data X and Y, given that there is some true relationship $Ex[y] = \beta_1 X + \beta_0$ is given by Gaussian distribution.

$$p(y|x, \theta) = \mathcal{N}(\beta^T x, \sigma^2)$$

This is a good assumption since the sum of arbitrary random variables tends toward a Gaussian by the central limit theorem.

# Maximum Likelihood Estimate (MLE)

- Assumes that the samples are from a specific distribution with some unknown parameter(s).
- **Likelihood** is the probability that the samples observed come from the given distribution. $L(\theta|X) = p(X|\theta)$
- We can estimate the parameter $\theta$ by maximizing the likelihood function.

# Maximum Likelihood Estimate (MLE)

Given a known distribution $f(x_i, \theta)$
Given the sample $x = \{x_1, \cdots, x_n\}$
The likelihood can be formulated

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

# Maximum Likelihood Estimate (MLE)

- MLE=The value of parameter $\theta$ that maximizes likelihood $L$ is the estimate
- This can be obtained by finding the derivative of the log-likelihood with respect to the parameter $\theta$ and equating the resulting formula to zero.
- Note: if there are multiple parameters, then you can maximize each one separately or attempt to find a multi-dimensional maxima.
  - since $\max L$ is the same as $\max \log L$ we can use

$$\ell = \log L = \log \prod_{i=1}^{n} f(x_i|\theta)$$

$$= \sum_{i=1}^{n} \log f(x_i|\theta)$$

# Theory of Linear Regression

- The PDF for Multivariate normal distribution(dimension=n)

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)^T\right) \tag{6}$$

- hence:

$$P(y = y_i|x = x_i) \sim \exp^{-\left(\frac{y_i-\beta_1 x_i-\beta_0}{2\sigma}\right)^2} \tag{7}$$

# Theory of Linear Regression

- The simplest method to find the parameters of the statistical model is based on the Maximum Likelihood Estimate:

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathcal{D}|\theta) \tag{8}$$

where $\mathcal{D}$, denotes the training set.

- Also, we can write(assuming training examples are i.i.d)

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(y_i|x_i,\theta) \sim \prod_{i=1}^{n} \exp^{-\left(\frac{y_i-\beta_1 x_i-\beta_0}{2\sigma}\right)^2} \tag{9}$$

# Theory of Linear Regression

- We use the multiplication rule of probability theory to find the total probability (likelihood) of the observed data set under this assumption
- Note that the linear model does not constrain $X$ to take on continuous values.

# Theory of Linear Regression

- However, we do not know, a priori, what the values of $\beta_1$ and $\beta_0$ are. We choose the values that make our observations maximally likely.
- Since the log of the probability is a monotonically increasing function, we can maximize the log of the probability, or minimize the negative square error:

$$-\text{Log}(P(\mathcal{D}|\theta)) \sim \underbrace{\sum_{i=1}^{n} \left( \frac{y_i - \beta_1 x_i - \beta_0}{2\sigma} \right)^2}_{\textit{Square error}} \tag{10}$$

We can do this by setting the derivative of the square error with respect to the coefficient $\beta_0$ and $\beta_1$ equal to zero.

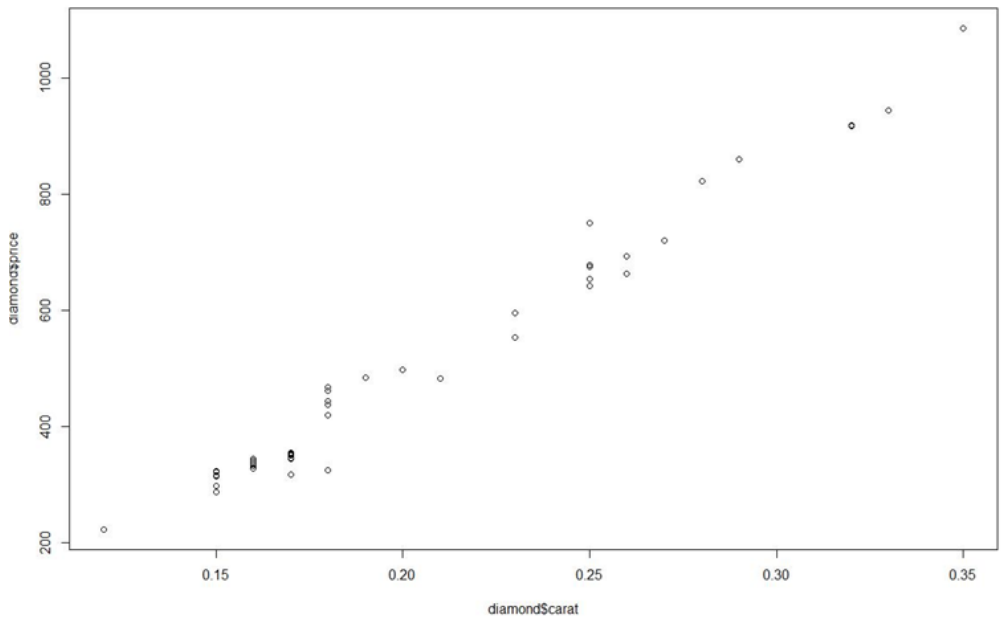
$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \left( \frac{y_i - \beta_1 x_i - \beta_0}{2\sigma} \right)^2 = 0 \tag{11}$$

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \left( \frac{y_i - \beta_1 x_i - \beta_0}{2\sigma} \right)^2 = 0 \tag{12}$$

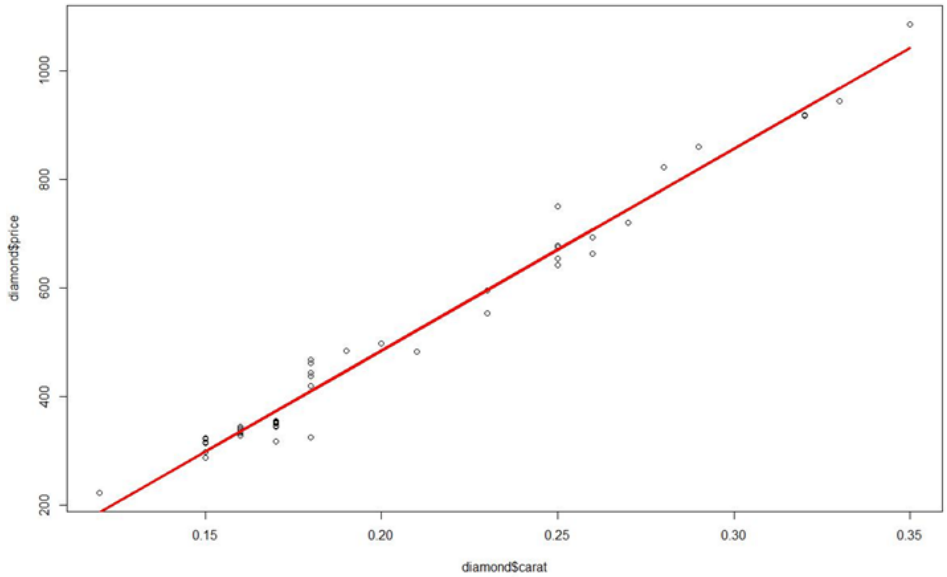
Note that the derivative is defined as long as $\beta_0$ and $\beta_1$ can take on continuous values.

# Simple Linear Regression

- Example: We have a data set on 48 diamond rings containing price in Singapore dollars and size of diamond in carats. The correlation between caret and price is 0.98. As you can see below, there is a linear relation in this data set.



# Simple Linear Regression




The equation of the line is expressed as follows

$$\text{Predicted Price} = 3721 * \text{Carat} - 259.6 \quad (13)$$

# Multivariable regression

- Now assume instead of having one variable as predictor we have a set of variables such as $\{X_0, X_1, \cdots, X_n\}$. Then the regression model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \tag{14}$$

  The we call a multivariable regression model.
- The method to find the best fitted multivariable regression is similar to simple regression. In other words, we need to minimize the distance between the predicted values and the actual values.

# Multivariable regression

- In this case we want to minimize

$$\sum_{i=1}^{n}(y_i - \beta^T x_i)^2 = [Y - X\beta]^T[Y - X\beta] \tag{15}$$

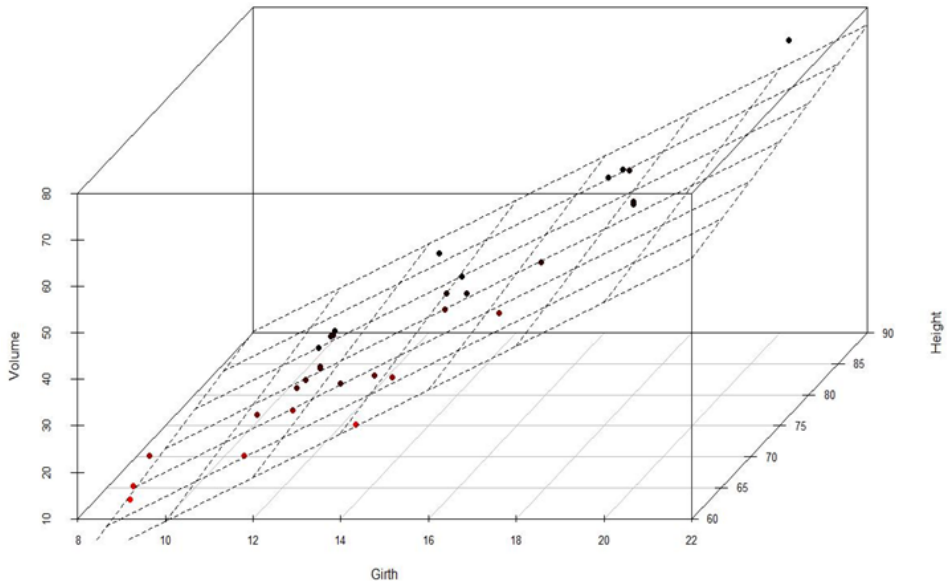
- The solution to such minimization is as follows

$$\beta = \left(X^T X\right)^{-1} X^T Y \tag{16}$$

- Example: We have a data set in which there are 3 variables describing different features of cherry trees:
  1. Girth: tree diameter in inches (denoted x1)
  2. Height: tree height in feet (x2).
  3. Volume: volume of the tree in cubic feet. (y)

  Let Y=volume and the predictors be Girth and Height.

# Multivariable regression




The regression equation for the plane is as follows:

$$\text{Predicted Volume} = 4.7 * \text{Girth} + 0.33 * \text{Height} - 57.98 \quad (17)$$

# Classification

- In machine learning, pattern recognition is the assignment of some sort of output value (or label) to a given input value (or instance), according to some specific algorithm.
- Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.
- Both Regression and Classification are aimed at finding a function $h$ which maps data $X$ to feature $y$. In regression, $y$ is a continuous variable.
- In classification, y is a discrete variable (categorical variable).
- In linear regression, data is modelled using a linear function, and unknown parameters are estimated from the data.

# Logistic Regression

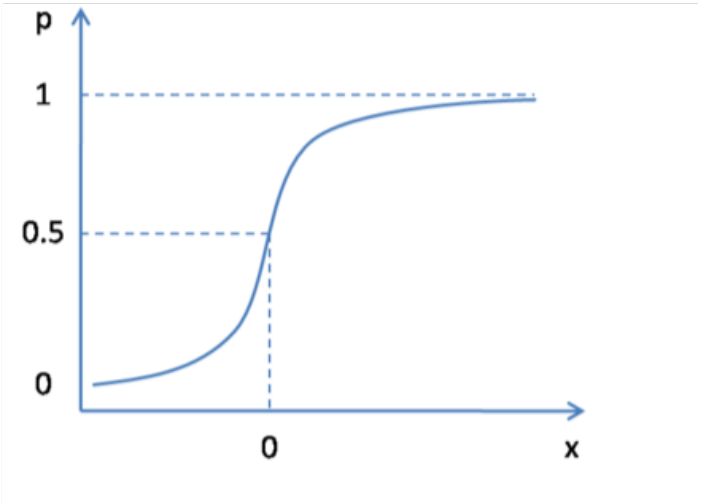
- In statistics, logistic regression (sometimes called the logistic model or logit model) is used for prediction of the probability of occurrence of an event by fitting data to a logit function logistic curve. (Walker, SH and Duncan, DB ,1967) It's a form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predictors.
- Addresses the same questions that discriminant function analysis and multiple regression do but with no distributional assumptions on the predictors (the predictors do not have to be normally distributed, linearly related or have equal variance in each group)
- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 1-positive and 0 negative.

# Logistic Regression

- We can define a function

$$p(y_i = 1|X = x_i) = \frac{e^{\beta_1 x_i + \beta_0}}{1 + e^{\beta_1 x_i + \beta_0}}$$

$$p(y_i|X = x_i) = (\frac{e^{\beta_1 x_i + \beta_0}}{1 + e^{\beta_1 x_i + \beta_0}})^{y_i}(1 - \frac{e^{\beta_1 x_i + \beta_0}}{1 + e^{\beta_1 x_i + \beta_0}})^{1-y_i}$$




- It looks similar to a step function, but we have relaxed it so that we have a smooth curve, and can therefore take the derivative.

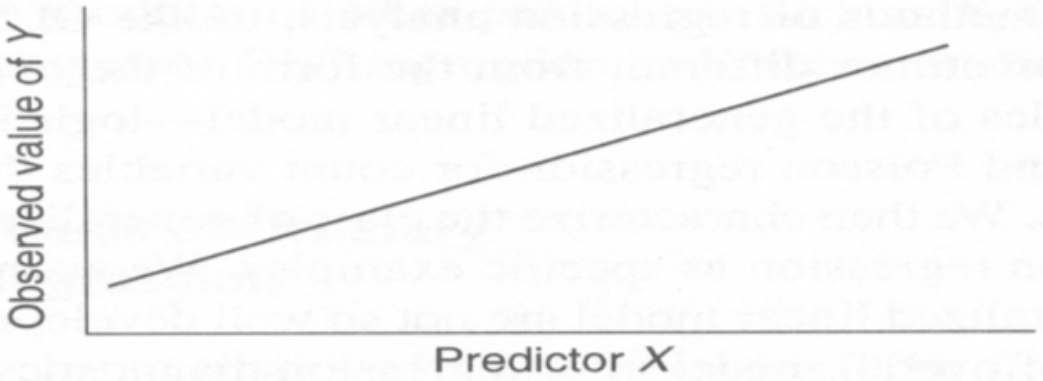
# Logistic Regression

Since the model is probabilistic.

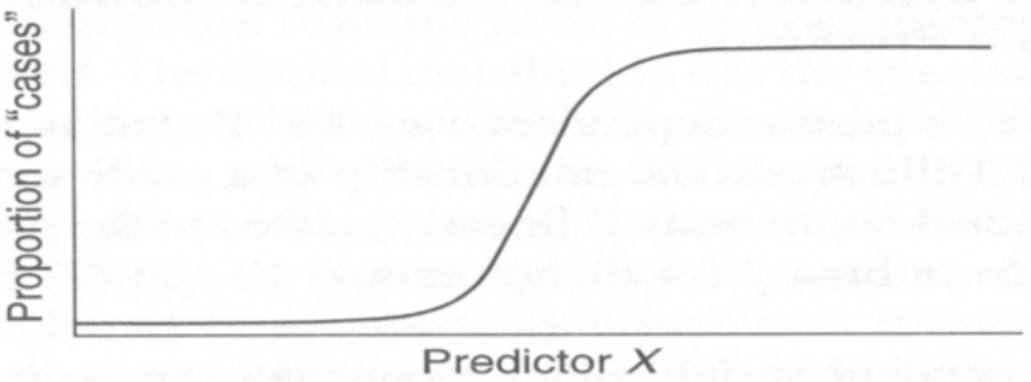
- This means that the output has to be able to provide us with probabilistic output
- The output has to be between 0 and 1.
- This allows to have two possible of outputs
  1. Categorical or binary
     - Threshold based output. If the probability is ¿ 0.5 then the output is 1
     - If the probability is ¡0.5, then the output is 0
  2. Probabilistic output. The outcome is produced in a shape of probabilities of it either being 0 or 1

# Logistic Regression

(A) For a continuous outcome variable $Y$, the numerical value of $Y$ at each value of $X$.




(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of $X$.



# Logistic Regression: Mathematical Theory

$$p(y_i|X = x_i) = (\frac{e^{\beta_1 x_i+\beta_0}}{1+e^{\beta_1 x_i+\beta_0}})^{y_i}(\frac{1}{1+e^{\beta_1 x_i+\beta_0}})^{1-y_i}$$

Using MLE, we can get the function parameters such that
$L(\beta) =$
$\sum y_i[(\beta_1 x_i+\beta_0)x_i - log(1+e^{\beta_1 x_i+\beta_0})]+(1-y_i)[-log(1+e^{\beta_1 x_i+\beta_0})]$

$$\mathcal{L}(\beta) = \sum y_i(\beta_1 x_i + \beta_0) - log(1 + e^{\beta_1 x_i+\beta_0})$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum y_i x_i^T - \frac{e^{\beta_1 x_i+\beta_0}}{1+e^{\beta_1 x_i+\beta_0}} x_i^T$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum (y_i - \frac{e^{\beta_1 x_i+\beta_0}}{1+e^{\beta_1 x_i+\beta_0}}) x_i^T$$

How to solve for $\beta$?

# Logistic Regression: Mathematical Theory

Let's use Newton- Raphson method

$$p(y_i|X = x_i) = (\frac{e^{\beta_1 x_i+\beta_0}}{1+e^{\beta_1 x_i+\beta_0}})^{y_i}(\frac{1}{1+e^{\beta_1 x_i+\beta_0}})^{1-y_i}$$

$$p_{(y_i=1)} = \frac{e^{\beta_1 x_i+\beta_0}}{1+e^{\beta_1 x_i+\beta_0}}$$

$$\beta^T x_i = \beta_1 x_i + \beta_0$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_i^n (y_i - p_{(y_i=1)}) x_i^T$$

$$\frac{\partial \mathcal{L}^2}{\partial \beta \partial \beta^T} = \sum_i^n -p_{(y_i=1)}(1 - p_{(y_i=1)}) x_i x_i^T$$

# Logistic Regression: Mathematical Theory

Let's use Newton- Raphson method

$$\beta^{new} = \beta^{old} - \frac{\frac{\partial \mathcal{L}}{\partial \beta}}{\frac{\partial \mathcal{L}^2}{\partial \beta \partial \beta^T}}$$