AI Course

Dr. Mürsel Taşgın

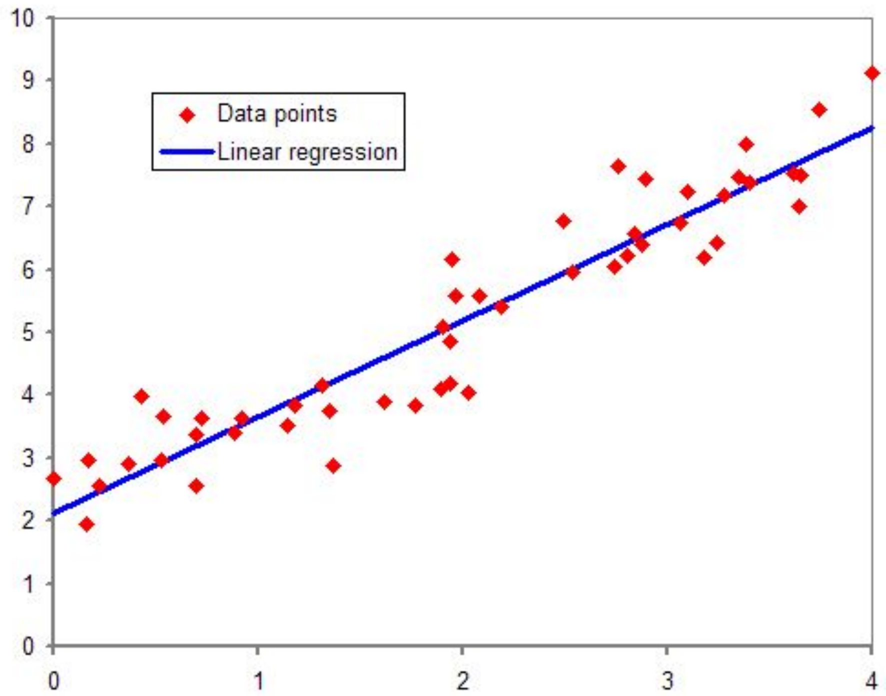
Regression

Regression- *Introduction*

- Regression analysis: Estimating the relationships between a *dependent variable (target)* and one or more *independent variables(features, inputs)*

- Regression is used to study the relationship between two (or more) variables

- Regression model
  - The **unknown parameters**, denoted as a scalar or a vector $\boldsymbol{\beta}$
  - The **independent variables**, observed in data, denoted as a vector $\boldsymbol{X_i}$
  - The **dependent variable**, observed in data, scalar $\boldsymbol{Y_i}$
  - The **error terms** *(residual)*, not directly observed in data, denoted as e_i



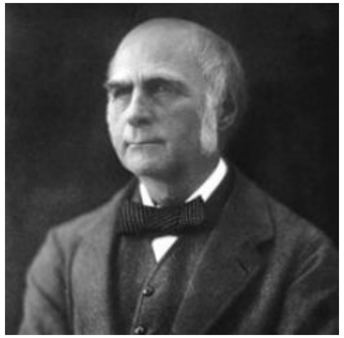
Regression- *Introduction*

History

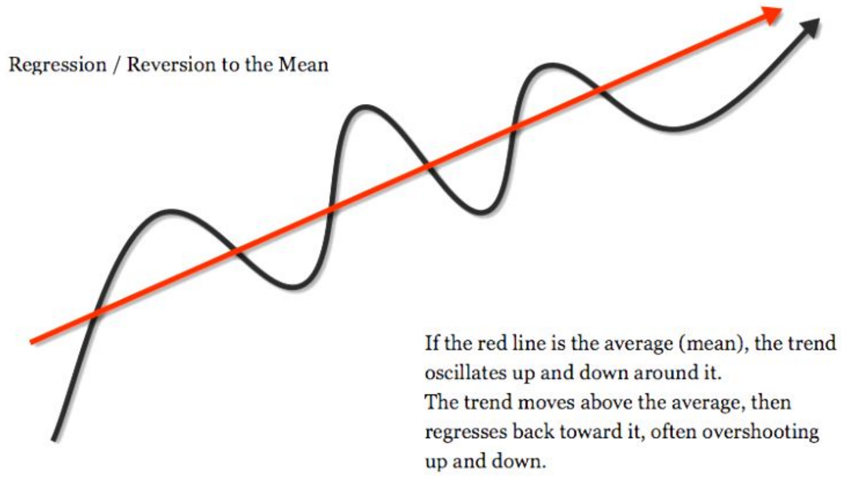
- The earliest form of regression was the *method of least squares*, which was published by Legendre in 1805, and by Gauss in 1809.

- The term "**regression**" was coined by Francis Galton in the 19th century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as *regression toward the mean*)

Carl Friedrich Gauss



Francis Galton



Linear Regression

Regression- *Linear Regression*

In linear regression, the model specification is that the dependent variable, y_i is a linear combination of parameters

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \dots, n.$$

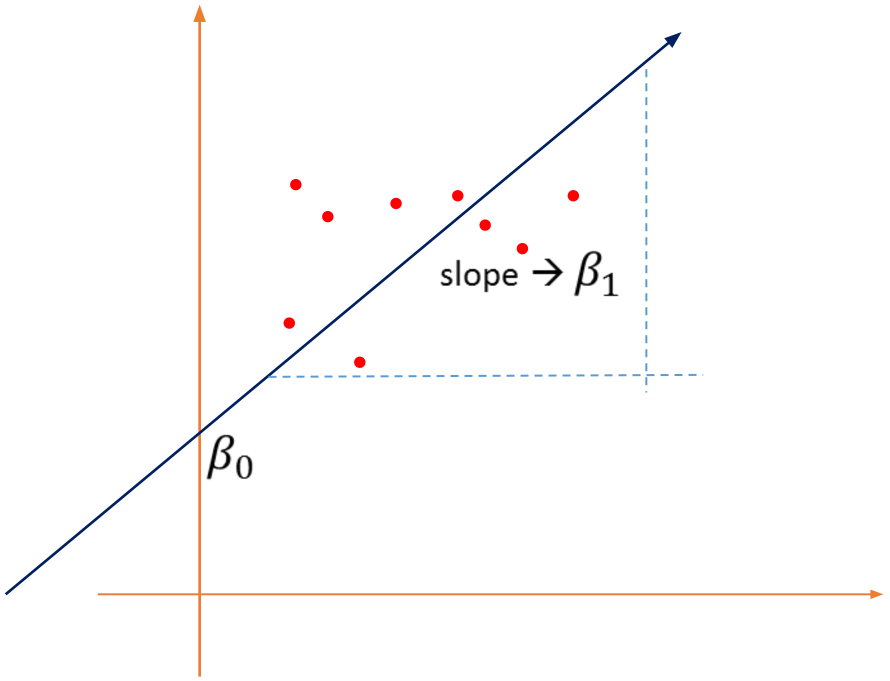
y_i the dependent variables

x_i the independent variables

β_0 is an intercept

β_1 is the coefficient

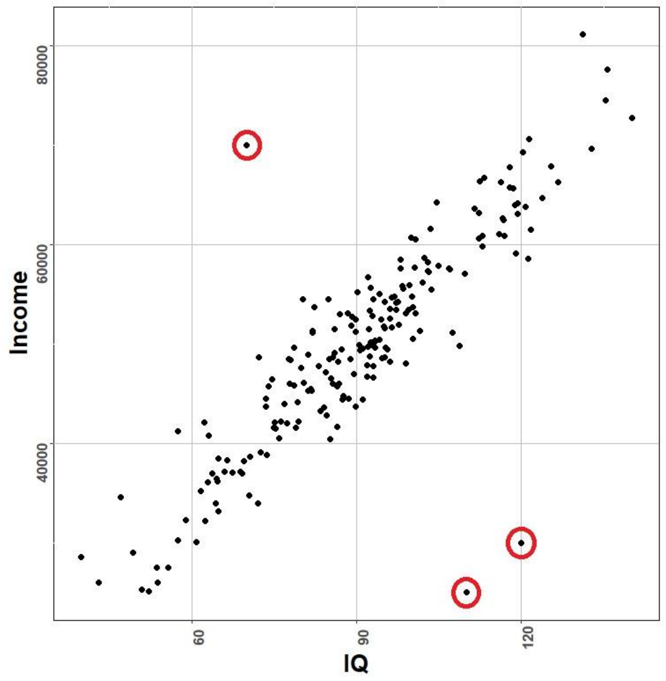
ϵ_i is an error term for each observation

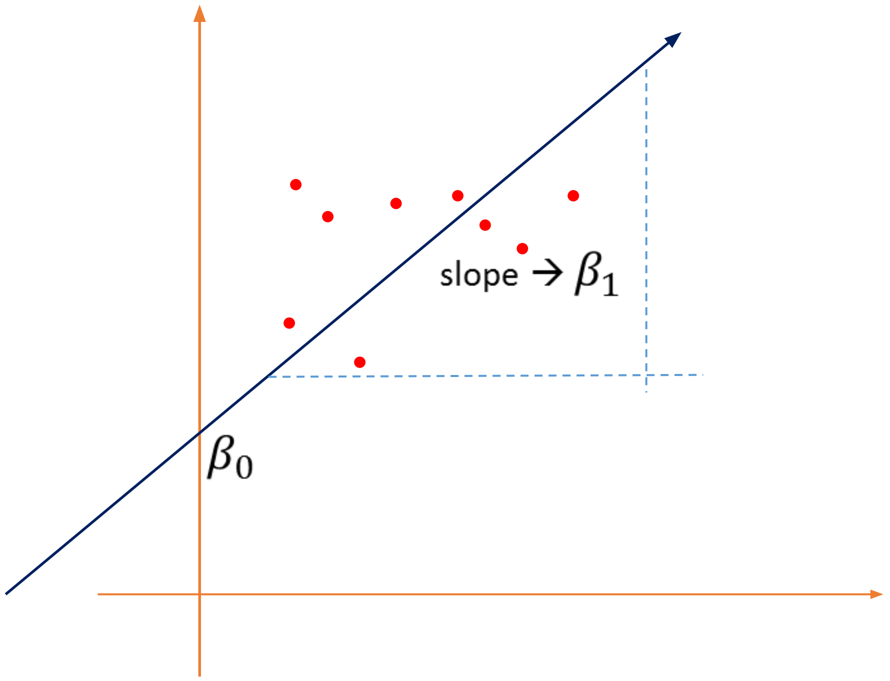


Regression- *Linear Regression*

Assumptions

- Relationship is linear
- The *y* values are distributed normally at each value of *x*
- The errors are normally distributed
- There are no clear outliers
- Observaions are independent





Regression- *Linear Regression*

Hypothesis testing

Regression tests the null hypothesis:

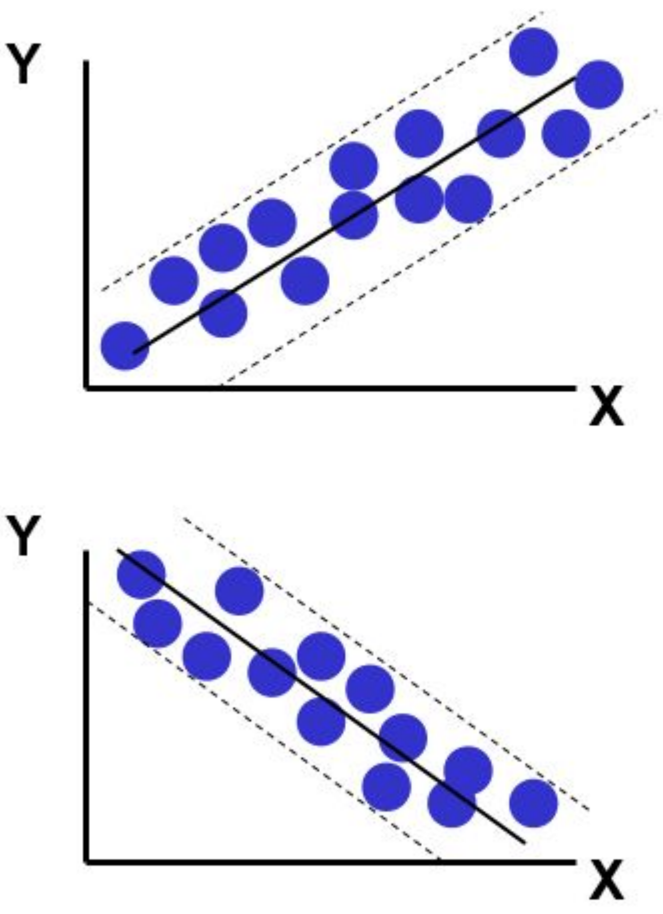
H_0 : There is no effect of X on Y, that is, β_1 = 0. Null hypothesis

versus the alternative hypothesis:

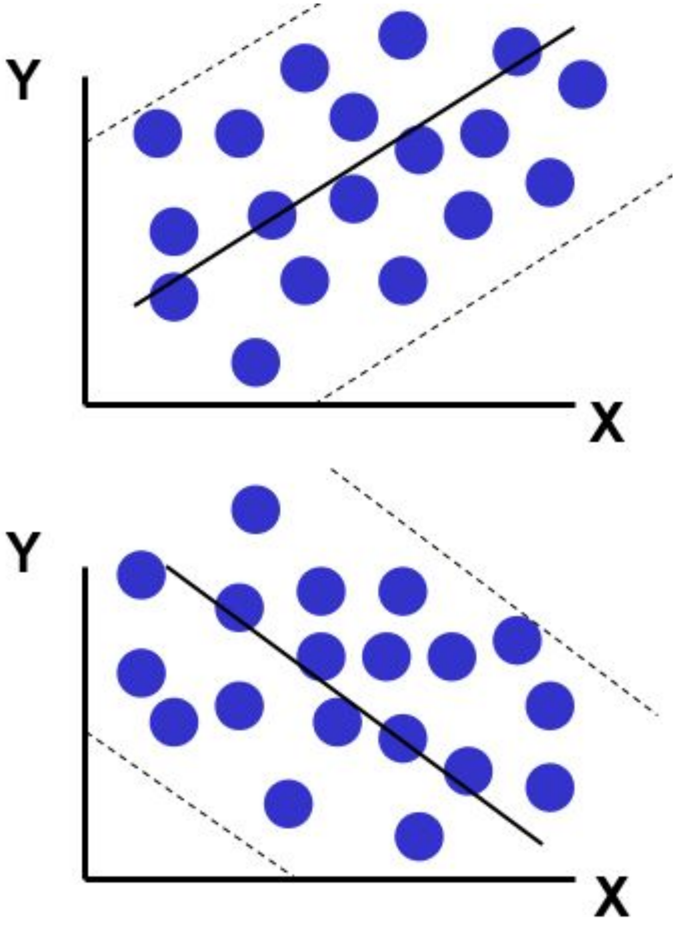
H_1 : There is an effect of X on Y, that is, β_1 is not 0.

If the null hypothesis is rejected, we reject the hypothesis that there is no relationship and hence we conclude that there is a significant relationship between X and Y.

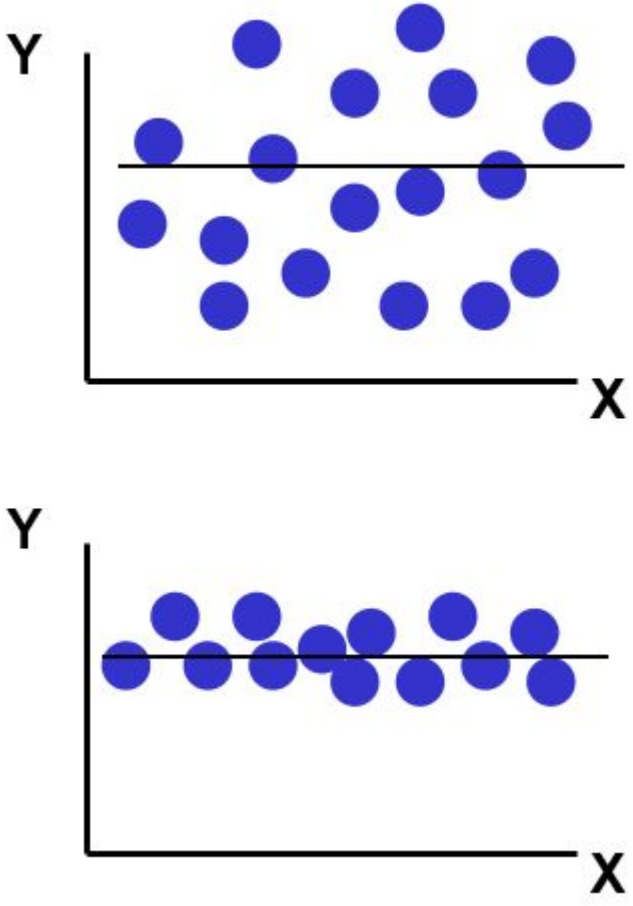
Regression- *Linear Regression*



Strong relationship



Weak relationship



No relationship

Regression- *Linear Regression*

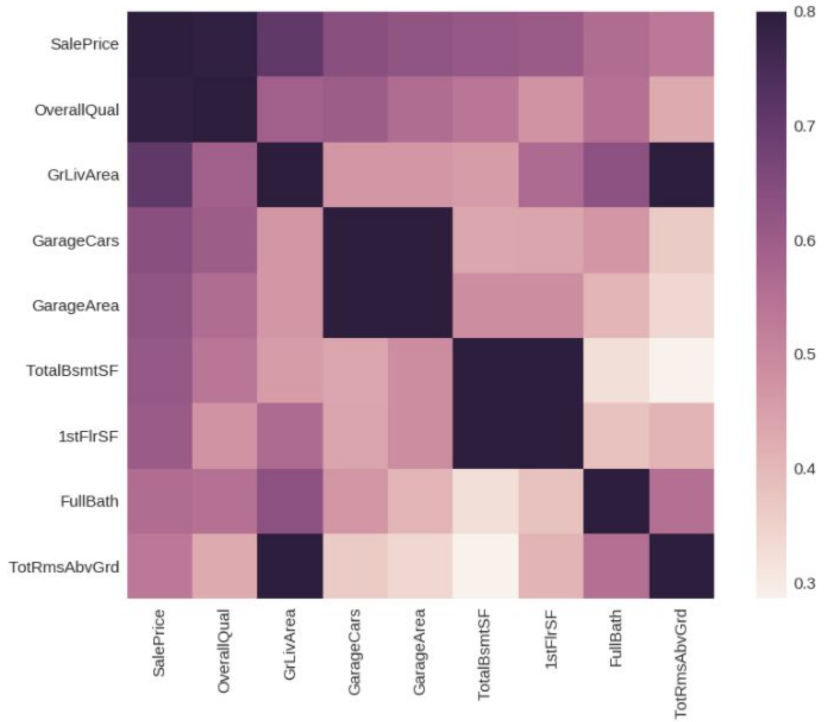
What can we do with regression analysis?

- Make predictions (based on available information)
- Estimate group means (for similar individuals)
- Measure effects (while controlling for other influences)
- Help evaluate/improve a model (of a relationship)

Regression- *Linear Regression*

Make a predictions

- House-price prediction
- Car maintenance cost prediction



Regression- *Linear Regression*

Measure and effect

A one-unit difference in an explanatory variable, *when everything else of relevance remains the same*, is typically associated with how large a difference in the dependent variable?

- Process: “Regress” the dependent variable onto *all* of the relevant explanatory variables (i.e., use the “most complete” model available).
- Answer: (coefficient of explanatory variable)

  ± (~2)·(standard error of coefficient)

- Example: Estimate the “pure” impact of 1,000 miles of driving during the year on annual maintenance costs.

Regression- *Linear Regression*

Estimate a group mean

For a group of similar individuals (i.e., individuals with the same values for several independent variables), estimate the mean value of the dependent variable.

- Process: "Regress" the dependent variable onto the given explanatory variables. Then "Predict." Fill in the values of the explanatory variables. Hit the "Predict" button.
- Answer: (prediction) ± (~2)·(standard error of estimated mean)

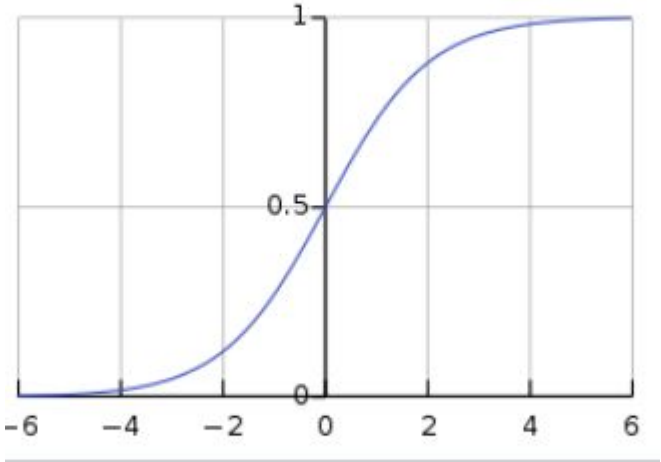
- Example: Estimate the mean annual maintenance cost of two-year-old Fords (*note the plural!*) in the fleet.

Logistic Regression

Regression- *Logistic Regression*

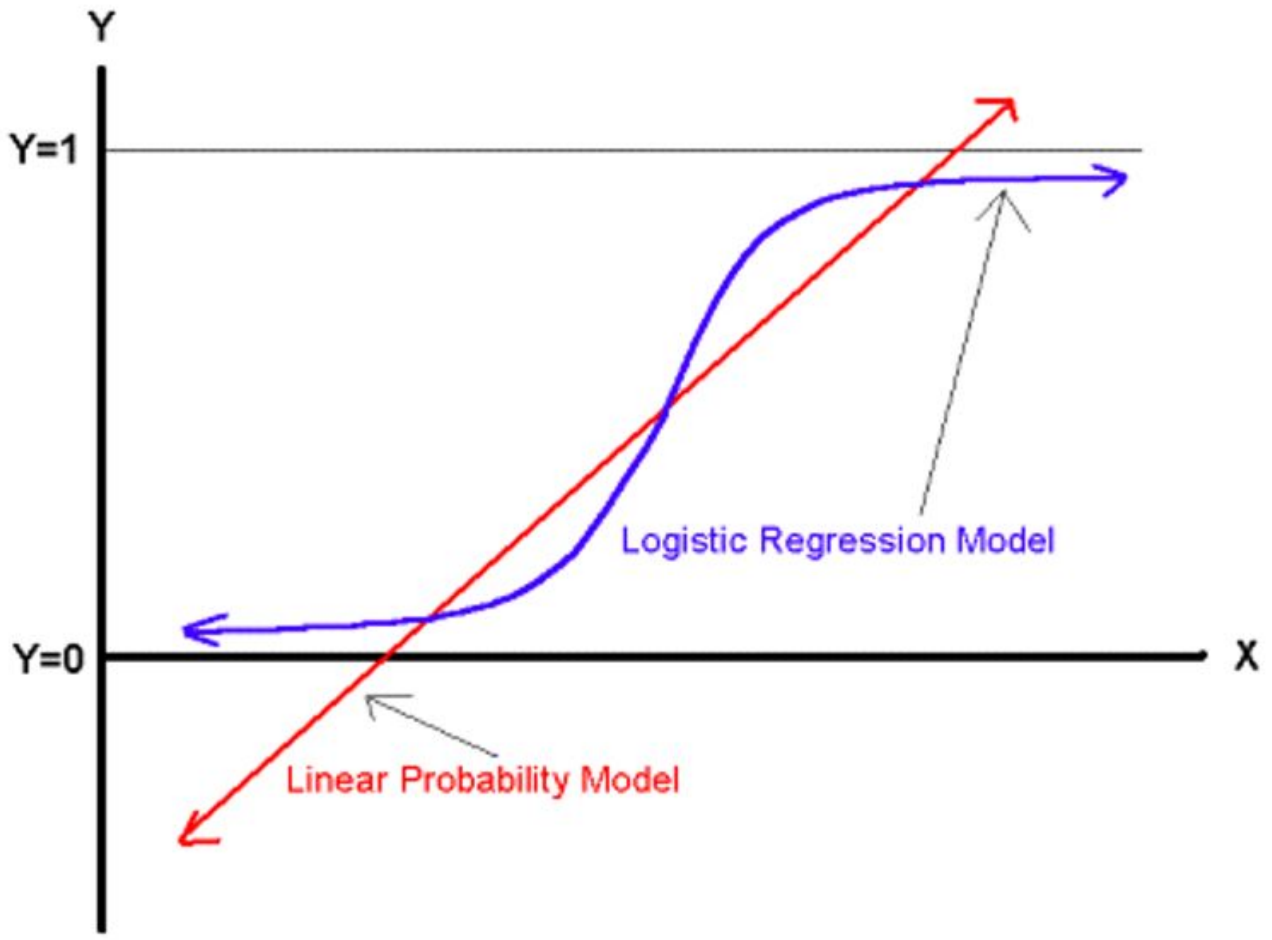
In statistics, the **logistic model** (or **logit model**) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.



Standart logistic function

Regression- *Logistic Regression*



Regression- *Logistic Regression*

The "logit" model :

$\ln[p/(1-p)] = \alpha + \beta X + e$

- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"

Regression- *Logistic Regression*

The "logit" model :

$$\ln[p/(1-p)] = \alpha + \beta X + e$$

- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

$$p = 1/[1 + \exp(-\alpha - \beta X)]$$

- if you let $\alpha + \beta X = 0$, then p = .50
- as $\alpha + \beta X$ gets really big, p approaches 1
- as $\alpha + \beta X$ gets really small, p approaches 0