Regression Analysis

Definition of Regression Analysis

Regression analysis is a statistical method used to estimate the relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as predictors or explanatory variables). The goal of regression analysis is to identify the strength and direction of the relationship between the variables, and to make predictions or draw inferences about the dependent variable based on the independent variables.

Types of Regression Analysis:

There are several types of regression analysis, including:

- Simple linear regression: a regression analysis that involves a single independent variable and a linear relationship between the independent and dependent variables.
- Multiple linear regression: a regression analysis that involves two or more independent variables and a linear relationship between the independent and dependent variables.
- Polynomial regression: a regression analysis that involves a curvilinear relationship between the independent and dependent variables.
- Logistic regression: a regression analysis that is used when the dependent variable is binary (e.g., yes/no, true/false) and the relationship between the independent and dependent variables is nonlinear.

Assumptions of Regression Analysis:

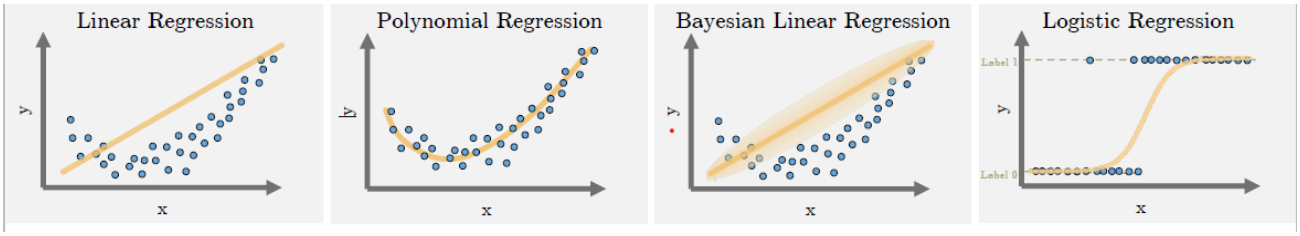
Regression analysis is based on several assumptions, including:

- Linearity: the relationship between the independent and dependent variables is linear.
- Independence: the observations are independent of each other.
- Homoscedasticity: the variance of the dependent variable is constant across different levels of the independent variable(s).
- Normality: the residuals (the differences between the predicted and actual values) are normally distributed.

Interpretation of Regression Analysis:

The output of a regression analysis typically includes several statistics, such as the coefficients, standard errors, and R-squared value. The coefficients represent the estimated change in the dependent variable for each unit change in the independent variable, while the standard errors reflect the degree of uncertainty in the coefficient estimates. The R-squared value indicates the proportion of variance in the dependent variable that is explained by the independent variables.

Visual Representation:



Summary:

	What does it fit?	Estimated function	Error Function
Linear	A line in n dimensions	$f_\beta^{linear}(x_i) = \beta_0 + \beta_1 x_i$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2$
Polynomial	A polynomial of order k	$f_\beta^{poly}(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2$
Bayesian Linear	Gaussian distribution for each point	$\mathcal{N}(f_\beta(x_i), \sigma^2)$	$\sum_i \|y_i - \mathcal{N}(f_\beta(x_i), \sigma^2)\|^2$
Ridge	Linear/polynomial	$f_\beta^{poly}(x_i)$ *or* $f_\beta^{linear}(x_i)$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{n} \beta_j^2$
LASSO	Linear/polynomial	$f_\beta^{poly}(x_i)$ *or* $f_\beta^{linear}(x_i)$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{n} \|\beta_j\|$
Logistic	Linear/polynomial with sigmoid	$\sigma(f_\beta(x_i))$	$min_\beta \sum_i -y_i log(\sigma(f_\beta(x_i))) - (1-y_i) log(1-\sigma(f_\beta(x_i)))$