Unit 5
Regression and Predictive Modeling

5.1 Empirical models, simple linear regression, MLE and least square estimator
5.2 Multiple linear regression, matrix approach to multiple linear regression, polynomial regression models, categorical regressors, indicator variables, selection of variables and model building
5.3 Logistic regression

Empirical Vs Theoretical Models

Empirical models:

An empirical models are models developed from observed data to describe or predict relationships between variables. They focus on identifying trends, correlations, and patterns, not necessarily grounded in theoretical physics or deterministic laws.

Examples: Model for Predicting house prices based on size or estimating demand for a product based on price. Model for predicting height of person based on age.

Theoretical Models:

A theoretical models are derived from established principles, laws, or theories to describe or predict a phenomenon. These models are based on prior scientific understanding or logical reasoning.

Example: Newton's Law of Universal Gravitation.

Use of Empirical Models:

1. **Predictive Analytics:** These models are used to forecast future events based on past data. For example, predicting stock prices, sales, or customer behavior.

2. **Regression Analysis:** Empirical models can be employed to understand and quantify relationships between variables in regression tasks.

3. **Risk Assessment:** These models can help in evaluating risks in finance, insurance, healthcare, etc.

Difference Between Regression and Classification

	Regression	Classification
Objective	The goal of regression is to predict a continuous numerical value.	The goal of classification is to predict a discrete label or class.
Type of Output	The output is a continuous value.	The output is a discrete class label.
Evaluation Metrics	Mean Squared Error (MSE), Root Mean Squared Error (RMSE),Mean Absolute Error (MAE), R-squared	Accuracy, Precision, Recall, F1-Score, Confusion Matrix
Examples	Predicting the temperature for the next day, Predicting the stock	Classifying emails as spam or not spam(binay classification),

	price of a company.	Categorizing reviews as positive, neutral, or negative (multiclass classification).
Algorithms Used	Simple linear regression, Multiple regression, Polynomial regression	Logistic Regression, Decision Trees Random Forests,

Role of Regression in Predictive Modeling

Regression plays a central role in predictive modeling by serving as the foundation for many empirical models. It enables the establishment of quantitative relationships between predictor variables (independent variables, X) and response variables (dependent variables,), making it one of the most widely used techniques for prediction and inference.

Types of Regression Models

Based on the type of functions used to represent the relationship between the dependent or output variable and independent variables, the regression models are categorized into four types. The regression models are,

1. Simple linear regression
2. Multiple regression
3. Polynomial regression

Simple Linear Regression Model:

Assume that there is only one independent variable x. If the relationship between x (independent variable) and y (dependent or output variable) is modeled by the relation,

$$y = a+bx$$

then the regression model is called a simple linear regression model.

Problem Definition:

Find a quadratic regression model for the following data:

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

Let the simple linear regression model be y = a + bx
Steps to find a and b,
First, find the mean and covariance.
Means of x and y are given by,

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{y} = \frac{1}{n}\sum y_i$$

The variance of x is given by,

$$\text{Var}\,(x) = \frac{1}{n-1}\sum (x_i - \bar{x}_i)^2$$

The covariance of x and y, denoted by Cov(x, y)is defined as,

$$\text{Cov}(x, y) = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$$

Now the values of a and b can be computed using the following formulas:

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

$$a = \bar{y} - b\bar{x}$$

First, find the mean of x and y,

$$n = 5$$

$$\bar{x} = \frac{1}{5}(1.0 + 2.0 + 3.0 + 4.0 + 5.0)$$

$$= 3.0$$

$$\bar{y} = \frac{1}{5}(1.00 + 2.00 + 1.30 + 3.75 + 2.25)$$

$$= 2.06$$

Next, find the Covariance between x and y,

$$\text{Cov}(x, y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Cov}(x, y) = \frac{1}{4}[(1.0 - 3.0)(1.00 - 2.06) + \cdots + (5.0 - 3.0)(2.25 - 2.06)]$$

$$= 1.0625$$

Now find the variance of x,

$$\text{Var}(x) = \frac{1}{n-1}\sum(x_i - \bar{x_i})^2$$

$$\text{Var}(x) = \frac{1}{4}[(1.0 - 3.0)^2 + \cdots + (5.0 - 3.0)^2]$$

$$= 2.5$$

Now, find the intercept and coefficients,

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{1.0625}{2.5}$$

$$= 0.425$$

$$a = 2.06 - 0.425 \times 3.0$$

$$= 0.785$$

Therefore, the linear regression model for the data is,

$$y = 0.785 + 0.425x$$

There are many ways to estimate the parameters given the study of the model for more than 100 years; nevertheless, there are two frameworks that are the most common. They are:

- ✓ Least Squares Estimation.
- ✓ Maximum Likelihood Estimation.

Both are optimization procedures that involve searching for different model parameters.

In short, Least squares Estimation is an approach to estimating the parameters of a model by seeking a set of parameters that results in the smallest squared error between the predictions of the model ($\bar{y}$) and the actual outputs (y), averaged over all examples in the dataset, so-called mean squared error.

Given a Simple Linear Regression,

$$\hat{Y} = \beta_0 + \beta_1 X$$

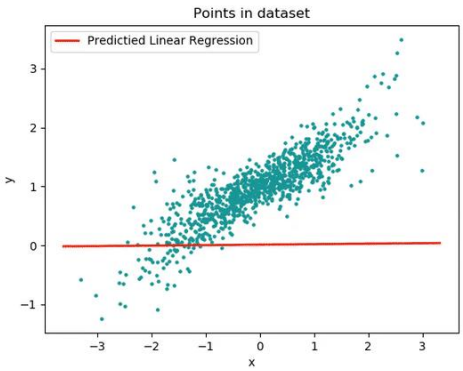
The least squares estimates of β_0 and β_1 are:

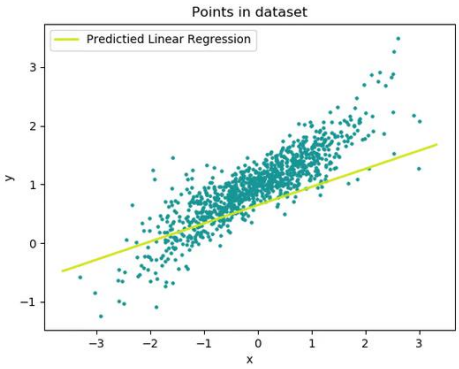
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

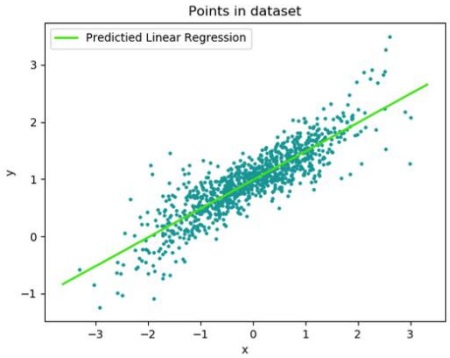
The classic derivation of the least squares estimates uses calculus to find the β0 and β1 parameter estimates that minimize the error sum of squares:

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Complete Derivation: Study
https://www.amherst.edu/system/files/media/1287/SLR_Leastsquares.pdf







We assume a model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i: Observed response variable
- x_i: Predictor variable
- β_0: Intercept
- β_1: Slope
- ϵ_i: Error term (assumed to have mean 0 and constant variance)

The residual for each observation is:

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

The residual sum of squares (RSS) is:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Our goal is to find β_0 and β_1 that minimize RSS.

To minimize RSS, set the partial derivatives with respect to β_0 and β_1 to 0:

(a) Partial derivative with respect to β_0:

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

Set it to 0:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

(b) Partial derivative with respect to β_1:

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

Set it to 0:

$$\sum_{i=1}^{n} x_i \left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

Substitute $\beta_0 = \bar{y} - \beta_1 \bar{x}$:

$$\sum_{i=1}^{n} x_i \left(y_i - (\bar{y} - \beta_1 \bar{x}) - \beta_1 x_i\right) = 0$$

Simplify:

$$\sum_{i=1}^{n} x_i \left(y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) - \beta_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) = 0$$

Rearrange:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

As,

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \bar{y} \sum_{i=1}^{n} x_i + n\bar{x}\bar{y}$$

Because of this we can write in terms of covariance.

While, Maximum Likelihood Estimation is a probabilistic method that seeks a set of parameters for the model that maximize a likelihood function.

(OPTIONAL)
Algorithm:

Step1: Define the Likelihood Function

The likelihood function represents the probability of observing the given data as a function of the parameters. For a dataset X=(x1,x2,⋯,xn), and a statistical model with a probability density function (PDF) or probability mass function (PMF) f(x | θ), the likelihood function L(θ) is the product of the individual likelihoods for each data point:

$$L(\theta) = P(X = x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Where θ is the parameter (or vector of parameters) of the model.

Step2: Log-Likelihood

To simplify the calculations, especially for products of small probabilities, the logarithm of the likelihood function is often used. This converts the product into a sum:

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

Step3: Maximize the Log-Likelihood

The goal is to find the parameter values $\hat{\theta}$ that maximize the log-likelihood. In practice, this is often done by taking the derivative of the log-likelihood function with respect to the parameter(s) and setting it equal to zero to find the maximum:

$$\frac{d}{d\theta}\ell(\theta) = 0$$

Step4: Solve for $\hat{\theta}$

Solve the equation to find the value of $\hat{\theta}$.

Example: MLE for Simple Linear Regression

1. The Linear Regression Model

The model is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ (i.e., the error term is normally distributed with mean 0 and variance σ^2).

Thus, the conditional probability density function of y_i given x_i is:

$$f(y_i|x_i, \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right).$$

2. Likelihood Function

For n observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the likelihood function is the joint probability of all y_i values:

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right).$$

3. Log-Likelihood Function

Taking the natural logarithm of the likelihood function simplifies it:

$$\ln L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

This is the **log-likelihood function**:

$$\ln L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

4. Partial Derivatives

Step 1: Differentiate with respect to β_0:

$$\frac{\partial}{\partial \beta_0}\ln L = \frac{\partial}{\partial \beta_0}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right].$$

The derivative of the squared term is:

$$\frac{\partial}{\partial \beta_0}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right]=\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i).$$

Set this equal to 0 to find the maximum likelihood estimate:

$$\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0.$$

Simplify:

$$n\beta_0+\beta_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i.$$

So:

$$\beta_0=\frac{1}{n}\sum_{i=1}^{n}y_i-\beta_1\frac{1}{n}\sum_{i=1}^{n}x_i.$$

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \quad \text{and} \quad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

Then:

$$\beta_0=\bar{y}-\beta_1\bar{x}.$$

Step 2: Differentiate with respect to β_1:

$$\frac{\partial}{\partial \beta_1}\ln L=\frac{\partial}{\partial \beta_1}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right].$$

The derivative is:

$$\frac{\partial}{\partial \beta_1}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right]=\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)(-x_i).$$

Set this equal to 0 to find the maximum likelihood estimate:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(x_i) = 0.$$

Simplify:

$$\sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Substitute $\beta_0 = \bar{y} - \beta_1 \bar{x}$:

$$\sum_{i=1}^{n} y_i x_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Simplify further to get:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Step 3: Differentiate with respect to σ^2:

$$\frac{\partial}{\partial \sigma^2} \ln L = \frac{\partial}{\partial \sigma^2} \left[-\frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \right].$$

The derivative is:

$$\frac{\partial}{\partial \sigma^2} \ln L = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Set this equal to 0:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Hence, β0, β1 and σ2 are calcualted.

Comparing Maximum Likelihood Estimation (MLE) and Least Squares Estimator (LSE) in Regression

1. Concept:

- ✓ MLE aims to find the parameter values that maximize the likelihood of observing the given data. It is based on the likelihood function, which describes the probability of the data given certain parameter values.
- ✓ LSE is a method of estimating the parameters by minimizing the sum of squared differences between the observed values and the values predicted by the model.

2. Assumptions:

· MLE assumes a probability distribution for the errors or residuals (e.g., normal, Poisson, etc.). In the case of linear regression, MLE assumes that the errors follow a normal distribution

· LSE assumes that the relationship between the dependent and independent variables is linear. It doesn't require a distribution for the errors but does assume that the errors are independent and identically distributed (i.i.d.) with constant variance(Homoscedasticity).

3. Objective Function:

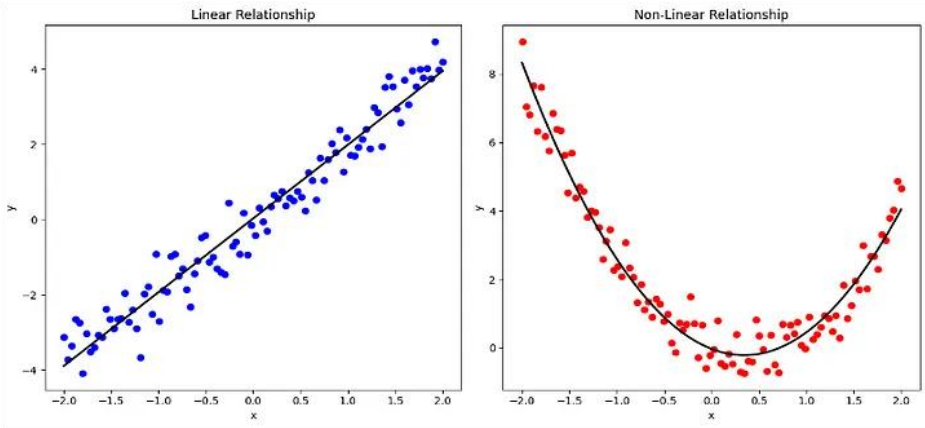
In MLE, the goal is to maximize the likelihood function. While in LSE, the goal is to minimize the sum of squared residuals:

4. General Applicability:

- ✓ MLE is more flexible and can be applied to a wide range of models beyond linear regression, such as logistic regression, Poisson regression, and more complex models.
- ✓ LSE is specific to linear models. It is commonly used when the relationship between the independent and dependent variables is linear.

FIVE KEY ASSUMPTIONS OF LINEAR REGRESSION ALGORITHM

1. Linear Relationship: It states that the dependent and independent variables should be linearly related.

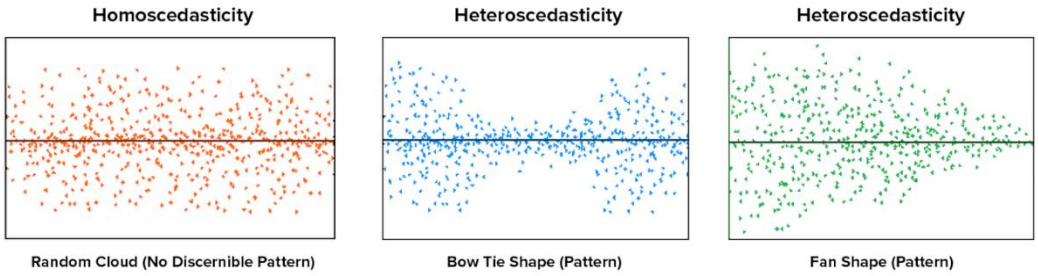


2. Normal Distribution of Residuals: The second assumption of linear regression is all the residuals or error terms should be normally distributed.

3. No Multicollinearity: The next assumption of linear regression is that there should be no multicollinearity in the given dataset. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated.

4. No Autocorrelation of Residuals: The residuals should be independent of each other. Autocorrelation occurs when residuals are correlated with previous or subsequent residuals.

5. Homoscedasticity: Homoscedasticity depicts a circumstance in which the residuals (that is, the "noise" or error terms in between the independent variables and the dependent variable) is the same across all values of the independent variables. Simply , residuals should have constant variance. If this condition is not followed, it is known as **heteroscedasticity**.
Heteroscedasticity leads to the unbalanced scatter of residuals or error terms. Generally, non-constant variation arises in the presence of outliers.



Some Applications of Simple Linear Regression

The following are some of the areas where Simple Linear Regression is used

1. Economics and Finance: Simple linear regression is employed in economics to analyse relationships between economic variables, such as the impact of interest rates on consumer spending or the relationship between inflation and unemployment.

2. Marketing and Sales: Businesses use simple linear regression for sales forecasting. By analysing historical sales data and factors like advertising expenditure or price changes, companies can make predictions about future sales and adjust their strategies accordingly.

3. Medical and Healthcare: Simple linear regression can be applied in healthcare to study the relationship between variables like patient age and medical expenses, drug dosage and treatment outcomes, or patient satisfaction and hospital wait times.

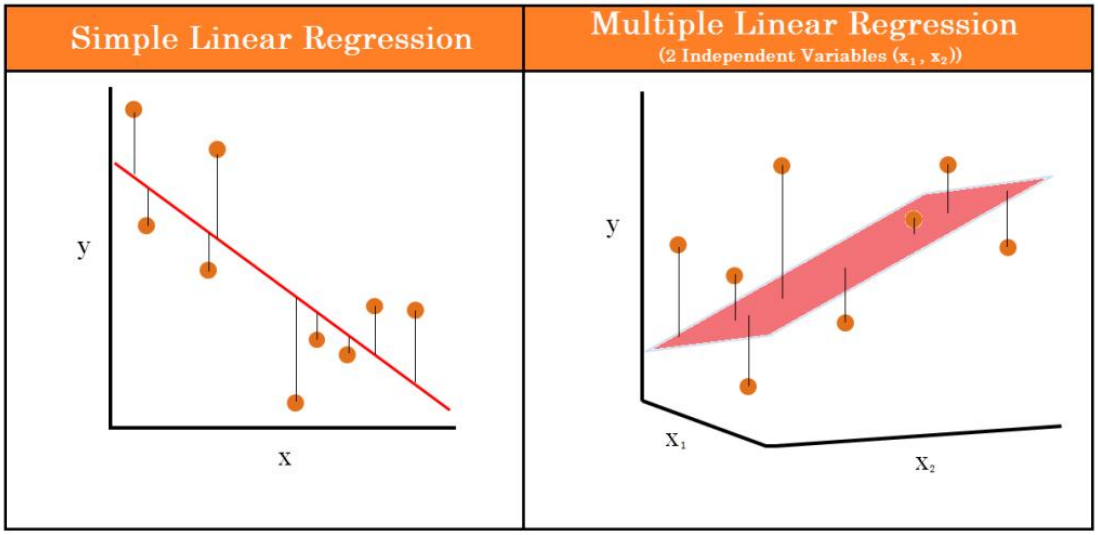
4. Sports Analytics: In sports analytics, simple linear regression can be used to analyze player performance metrics (e.g., batting average in baseball or shooting percentage in basketball) and their relationship with factors like training intensity, player fatigue, or coaching strategies.

5. Energy and Utilities: Energy companies can use simple linear regression to predict energy consumption based on historical data and weather conditions. This helps in resource planning and optimizing energy distribution.

A lot of areas are still there where linear relationships between variables persist.

Multiple Linear Regression:

Multiple Linear Regression (MLR) is a statistical technique used to model the relationship between one dependent variable and two or more independent variables. It extends simple linear regression by fitting a line (or hyperplane) in higher dimensions to describe the relationship.



The multiple regression of two variables x1 and x2 is given as follows:

$$y = f(x_1, x_2)$$

$$y = a_0 + a_1x_1 + a_2x_2$$

In general, this is given for 'n' independent variables as:

$$y = f(x_1, x_2, \ldots . x_n)$$

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots \ldots + a_nx_n + \varepsilon$$

Here, $x_1, x_2, \ldots . x_n$ are predictor variables, y is the dependent variable, $(a_0, a_1, a_2, \ldots . a_n)$ are the coefficients of the regression equation and ε is the error term.

The Multiple Linear Regression Model with 2 independent variables is written as follows:

$$Y=a+b_1X_1+b_2X_2+\varepsilon$$

where,
Y = The variable needs to be predicted (dependent variable)
X = The variable used to predict Y (independent variable)

a = The intercept
b = The slope
ε= The regression residual

Regression of two independent variables can be predicted by using the below formulas such as Intercepts (a), Regression Coefficients (b1, b2)

1. **Intercept (a):**

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

2. **Regression Coefficient b_1:**

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

3. **Regression Coefficient b_2:**

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

4. **Sum of squares and cross-products:**

- $\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)^2}{N}$
- $\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)^2}{N}$
- $\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$
- $\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$
- $\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$

Where:

- $\bar{Y}$, $\bar{X}_1$, and $\bar{X}_2$ are the means of Y, X_1, and X_2, respectively.
- N is the number of data points.

Example:

Subject	Y	X1	X2
1	-3.7	3	8
2	3.5	4	5
3	2.5	5	7
4	11.5	6	3
5	5.7	2	1
6	?	3	2

Step 1

First, calculate all the values required in the above formulae.

Subject	Y	X_1	X_2	X_1X_2	X_1X_1	X_2X_2	X_1Y	X_2Y
1	-3.7	3	8	24	9	64	-11.1	-29.6
2	3.5	4	5	20	16	25	14	17.5
3	2.5	5	7	35	25	49	12.5	17.5
4	11.5	6	3	18	36	9	69	34.5
5	5.7	2	1	2	4	1	11.4	5.7
SUM	**19.5**	**20**	**24**	**99**	**90**	**148**	**95.8**	**45.6**

Step 2

Then put these values into the above-mentioned formulae to get the exact predictable values to calculate Regression Coefficients b1 and b2

$$\sum x_1^2 = \sum X_1X_1 - \frac{(\sum X_1)(\sum X_1)}{N} = 90 - \frac{20 \times 20}{5} = 10$$

$$\sum x_2^2 = \sum X_2X_2 - \frac{(\sum X_2)(\sum X_2)}{N} = 148 - \frac{24 \times 24}{5} = 32.8$$

$$\sum x_1y = \sum X_1Y - \frac{(\sum X_1)(\sum Y)}{N} = 95.8 - \frac{20 \times 19.5}{5} = 17.8$$

$$\sum x_2y = \sum X_2Y - \frac{(\sum X_2)(\sum Y)}{N} = 45.6 - \frac{24 \times 19.5}{5} = -48$$

$$\sum x_1x_2 = \sum X_1X_2 - \frac{(\sum X_1)(\sum X_2)}{N} = 99 - \frac{20 \times 24}{5} = 3$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1y) - (\sum x_1x_2)(\sum x_2y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$b_1 = \frac{(32.8 \times 17.8) - (3 \times (-48)}{(10 \times 32.8) - (3)^2} = 2.2816$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2y) - (\sum x_1x_2)(\sum x_1y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$b_1 = \frac{(10 \times (-48)) - (3 \times 17.8)}{(10 \times 32.8) - (3)^2} = -1.672$$

Step 3:

Calculate the value of Intercept a

$$a = \overline{Y} - b_1(\overline{X}_1) - b_2(\overline{X}_2) = \frac{19.5}{5} - \frac{2.2816 \times 20}{5} - \frac{(-1.672 \times 24)}{5} = 2.796$$

Step 4:

The final Regression Equation or Model looks as follows:

$$Y = 2.796 + 2.28x_1 - 1.67x_2$$

Therefore, for given **x1= 3** and **x2 = 2**, the value of **Y = ?** calculated as follows:

$$Y = 2.796 + (2.28 \times 3) - (1.67 \times 2)$$

$$Y = 6.296$$

Assumptions and Diagnostics in MLR

Assumptions:

- ✓ Linearity: Relationship between predictors and response is linear.
- ✓ Independence: Observations are independent.
- ✓ Homoscedasticity: Constant variance of errors.
- ✓ Normality of Errors: Errors (ϵ) are normally distributed.
- ✓ No Multicollinearity: Predictors are not highly correlated.

Diagnostics in multiple linear regression refer to the techniques and methods used to assess the validity and reliability of the regression model. These diagnostics help identify whether the model meets the underlying assumptions and whether the results are interpretable and robust.

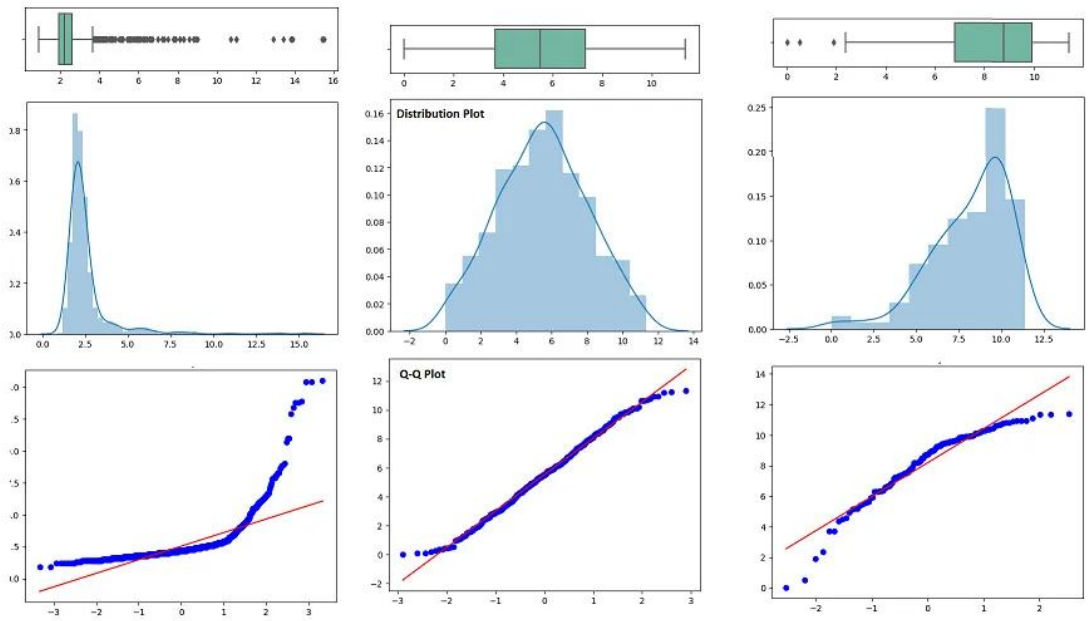
- ✓ Residual Analysis

  Residuals ($e_i = Y_i - \hat{Y}_i$) are the differences between observed and predicted values. Examining residuals can reveal violations of assumptions.

  Linearity check: Plot Residuals vs. Predicted Values. It should show no clear patterns(curves) when plotted against predicted values.

  Homoscedasticity Check: Residuals should have constant variance (no "funnel" shape in residual vs. fitted plot)

  Normality of Residuals: Residuals should be approximately normally distributed. Also, Use Q-Q plots (Quantile-Quantile Plot)



✓ Variance Inflation Factor (VIF)

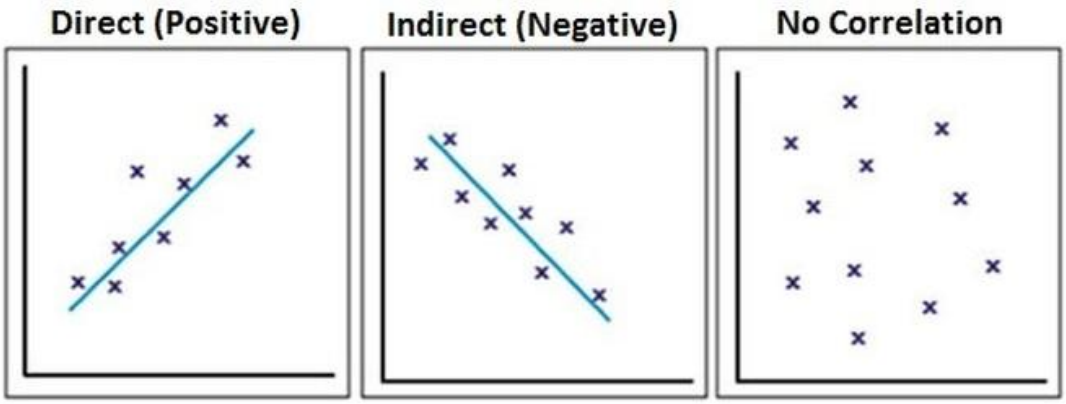
VIF helps detect multicollinearity, which occurs when two or more predictors are highly correlated with each other. High multicollinearity can lead to unstable estimates of regression coefficients. VIF > 10 indicates severe multicollinearity.

✓ Outlier Detection

Use standardized residuals (z-scores) to identify outliers. Points with standardized residuals >3 or <−3 are potential outliers.

✓ Autocorrelation test(Durbin-Watson Test)

Values close to 2 indicate no autocorrelation. Values < 2 suggest positive autocorrelation; > 2 indicate negative autocorrelation.



Matrix approach to multiple linear regression:

Consider a dataset with two independent variables X1 and X2, and one dependent variable Y as follows:

X1	X2	Y2
2	3	6
4	4	10
5	7	12
6	8	14
8	10	20

We need to solve for the coefficients b_0, b_1, and b_2 using the matrix method.
As, the equation for multiple linear regression is:

$$Y=b_0+b_1X_1+b_2X_2+\epsilon$$

Matrix Formulation:

The multiple linear regression model can be written as:

$$Y = X\beta$$

Where:

- Y is the vector of dependent variable observations,
- X is the matrix of independent variables (including the intercept term),
- β is the vector of coefficients we need to find, which includes the intercept and regression coefficients.

The normal equation to solve for β is:

$$\beta = (X^T X)^{-1} X^T Y$$

Where, β holds 3 values b0,b1,b2.

Step 1: Create the X matrix (including the intercept term) and the Y vector.

$$X = \begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ 1 & X_1^{(2)} & X_2^{(2)} \\ 1 & X_1^{(3)} & X_2^{(3)} \\ 1 & X_1^{(4)} & X_2^{(4)} \\ 1 & X_1^{(5)} & X_2^{(5)} \end{bmatrix}$$

$$Y = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \\ Y^{(3)} \\ Y^{(4)} \\ Y^{(5)} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ 1 & X_1^{(2)} & X_2^{(2)} \\ 1 & X_1^{(3)} & X_2^{(3)} \\ 1 & X_1^{(4)} & X_2^{(4)} \\ 1 & X_1^{(5)} & X_2^{(5)} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 4 & 5 \\ 1 & 5 & 7 \\ 1 & 6 & 8 \\ 1 & 8 & 10 \end{bmatrix}$$

$$Y = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \\ Y^{(3)} \\ Y^{(4)} \\ Y^{(5)} \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 12 \\ 14 \\ 20 \end{bmatrix}$$

2. **Step 2: Calculate $X^T X$.** First, calculate the transpose of X, X^T:

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 5 & 6 & 8 \\ 3 & 5 & 7 & 8 & 10 \end{bmatrix}$$

Then, compute $X^T X$:

$$X^T X = \begin{bmatrix} 5 & 25 & 33 \\ 25 & 174 & 224 \\ 33 & 224 & 290 \end{bmatrix}$$

3. **Step 3: Calculate $X^T Y$.** Multiply X^T with Y:

$$X^T Y = \begin{bmatrix} 6 + 10 + 12 + 14 + 20 \\ 2(6) + 4(10) + 5(12) + 6(14) + 8(20) \\ 3(6) + 5(10) + 7(12) + 8(14) + 10(20) \end{bmatrix} = \begin{bmatrix} 62 \\ 376 \\ 472 \end{bmatrix}$$

4. **Step 4: Compute the inverse of $X^T X$.** The inverse of a 3x3 matrix can be computed using standard matrix methods. For simplicity, I'll show the result after the matrix inversion:

$$(X^T X)^{-1} = \begin{bmatrix} 1.66 & -0.39 & 0.12 \\ -0.39 & 0.13 & -0.07 \\ 0.12 & -0.07 & 0.05 \end{bmatrix}$$

5. **Step 5: Solve for β.** Now, we can solve for $\beta = (X^T X)^{-1} X^T Y$:

$$\beta = \begin{bmatrix} 1.66 & -0.39 & 0.12 \\ -0.39 & 0.13 & -0.07 \\ 0.12 & -0.07 & 0.05 \end{bmatrix} \begin{bmatrix} 62 \\ 376 \\ 472 \end{bmatrix}$$

After multiplying, we get the regression coefficients:

$$\beta = \begin{bmatrix} 2.67 \\ 1.28 \\ 0.84 \end{bmatrix}$$

Final Answer:

The estimated regression equation is:

$$Y = 2.67 + 1.28X_1 + 0.84X_2$$

This means that:

- The intercept $b_0 = 2.67$,
- The coefficient for X_1 (b_1) is 1.28, and
- The coefficient for X_2 (b_2) is 0.84.

This is the multiple linear regression model fitted to the data using the matrix method.

Derivation of regression coefficients using matrix algebra.

Recall the model equation we use in linear regression:

$$Y_i = \beta_0 + \beta_1(X_{1_i}) + \beta_2(X_{2_i}) + \ldots + \beta_k(X_{k_i}) + \epsilon_i$$

can be written as,

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1(X_{1_1}) + \beta_2(X_{2_1}) + \ldots + \beta_k(X_{k_1}) + \epsilon_1 \\
Y_2 &= \beta_0 + \beta_1(X_{1_2}) + \beta_2(X_{2_2}) + \ldots + \beta_k(X_{k_2}) + \epsilon_2 \\
Y_3 &= \beta_0 + \beta_1(X_{1_3}) + \beta_2(X_{2_3}) + \ldots + \beta_k(X_{k_3}) + \epsilon_3 \\
&\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
Y_n &= \beta_0 + \beta_1(X_{1_n}) + \beta_2(X_{2_n}) + \ldots + \beta_k(X_{k_n}) + \epsilon_n
\end{aligned}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0(1) + \beta_1(X_{1_1}) + \beta_2(X_{2_1}) + \ldots + \beta_k(X_{k_1}) \\ \beta_0(1) + \beta_1(X_{1_2}) + \beta_2(X_{2_2}) + \ldots + \beta_k(X_{k_2}) \\ \beta_0(1) + \beta_1(X_{1_3}) + \beta_2(X_{2_3}) + \ldots + \beta_k(X_{k_3}) \\ \vdots \\ \beta_0(1) + \beta_1(X_{1_n}) + \beta_2(X_{2_n}) + \ldots + \beta_k(X_{k_n}) \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1_1} & X_{2_1} & \dots & X_{k_1} \\ 1 & X_{1_2} & X_{2_2} & \dots & X_{k_2} \\ 1 & X_{1_3} & X_{2_3} & \dots & X_{k_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1_n} & X_{2_n} & \dots & X_{k_n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We can write this as,

$$y = X\,b + e$$

we want to find the coefficient values that produce the smallest sum of squared residuals. To do this, we first re-write the regression equation to isolate the error vector:

$$\mathbf{e = y - X\,b}$$

The sum of squared residual can be expressed in matrix notation as e^Te.

$$\mathbf{e^Te = (y - X\,b)^T(y - X\,b)}$$

Using the rules of transposes and expanding the right-hand side, we get,

$$\mathbf{y^Ty - b^TX^Ty - y^TX\,b + b^TX^TX\,b}$$

Now, Each of these terms is a 1×1 matrix, which implies that each term is equal to its transpose. We will re-write the third term $y^TX\,b$ as its transpose b^TX^Ty. Re-writing, we get:

$$\mathbf{y^Ty - b^TX^Ty - b^TX^Ty + b^TX^TX\,b}$$

It becomes:

$$\mathbf{y^Ty - 2\,b^TX^Ty + b^TX^TX\,b}$$

To find the values for the elements in b that minimize the equation, we differentiate this expression with respect to b.

$$\frac{\delta}{\delta \mathbf{b}}(\mathbf{y^Ty - 2\,b^TX^Ty + b^TX^TX\,b})$$

we get,

$$\mathbf{-2\,X^Ty + 2\,X^TX\,b}$$

We set this equal to zero and solve for b.

$$\mathbf{-2\,X^Ty + 2\,X^TX\,b = 0}$$

$$\mathbf{2\,X^TX\,b = 2\,X^Ty}$$

$$\mathbf{X^T X\, b = X^T y}$$

To isolate b we pre-multiply both sides of the equation by $(\mathbf{X^T X})^{-1}$.

$$(\mathbf{X^T X})^{-1}\,(\mathbf{X^T X})\,\mathbf{b} = (\mathbf{X^T X})^{-1}\,\mathbf{X^T y}$$

$$\mathrm{I}\,\mathbf{b} = (\mathbf{X^T X})^{-1}\,\mathbf{X^T y}$$

Now, The vector of regression coefficients can be obtain from:

$$\mathbf{b} = (\mathbf{X^T X})^{-1}\,\mathbf{X^T y}$$

This implies that the vector of regression coefficients can be obtained directly through manipulation of the design matrix and the vector of outcomes. In other words, the OLS coefficients is a direct function of the data.

Reference: Fox, J. (2009). A mathematical primer for social statistics

Computational Advantages of the Matrix Approach

1. The matrix form allows for a more compact and efficient representation of linear regression models. Instead of solving equations for each individual variable, we can represent all the equations in a single matrix equation.

2. Matrix operations such as matrix multiplication and inversion are optimized and implemented in highly efficient libraries (e.g., NumPy, MATLAB, etc.).This allows for faster computation.

3. The matrix approach scales well to large datasets with many variables.

Polynomial Regression Model:

Polynomial regression is a type of regression analysis where the relationship between the independent variable X and the dependent variable Y is modeled as an nth-degree polynomial. It is used when the data exhibits a nonlinear relationship that can be better captured by higher-degree polynomials rather than just a straight line.

General Polynomial Regression Model:

The general form of the polynomial regression model is:

$$Y = b_0 + b_1X + b_2X^2 + b_3X^3 + \cdots + b_nX^n + \epsilon$$

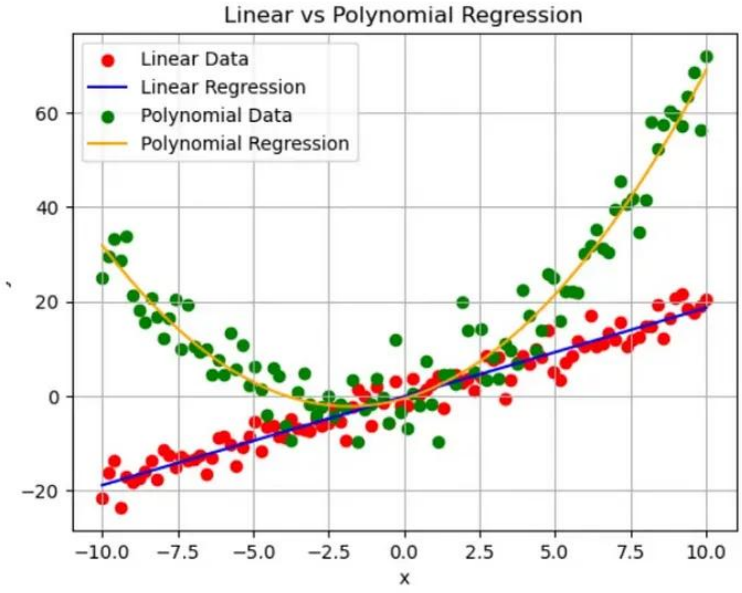
Where:

- Y is the dependent variable (response),
- X is the independent variable (predictor),
- $b_0, b_1, b_2, \ldots, b_n$ are the coefficients (parameters) to be estimated,
- n is the degree of the polynomial,
- ϵ is the error term (random noise).

Why Use Polynomial Regression?

✓ Handles Non-Linear Relationships: Captures patterns that a straight line can't represent.

✓ Flexible Modeling: By increasing the degree of the polynomial, you can adjust the curve to better fit the data.



The problem of non-linear regression can be solved by two methods:

1. Transformation of non-linear data to linear data, so the linear regression model can handle the data.

Consider a non-linear equation,

$$y = ae^{bx}$$

1. **Take the natural logarithm (ln) of both sides:**

$$\ln(y) = \ln(ae^{bx})$$

2. **Use logarithm properties** to simplify:

$$\ln(y) = \ln(a) + \ln(e^{bx})$$

$$\ln(y) = \ln(a) + bx$$

3. **Let $Y = \ln(y)$ and $A = \ln(a)$, then the equation becomes:**

$$Y = A + bx$$

Hence, we convert it into linear form.

For example: https://www.youtube.com/watch?v=lA-LVn5Rczo&list=PLRfhFQlD9tkIlUCy_luEciM3YTjK4TIVZ&index=2

2. Using Polynomial regression algorithm.

Example:

Conside the data points:

X	Y
1	1
2	4
3	9
4	15

We are given the quadratic regression equation: $y=a_0+a_1x+a_2x^2$

The system of equations to the above equation are:

1. $\sum y = n \cdot a_0 + a_1 \cdot \sum x + a_2 \cdot \sum x^2$
2. $\sum x \cdot y = a_0 \cdot \sum x + a_1 \cdot \sum x^2 + a_2 \cdot \sum x^3$
3. $\sum x^2 \cdot y = a_0 \cdot \sum x^2 + a_1 \cdot \sum x^3 + a_2 \cdot \sum x^4$

Let's first construct the table for all these values.

x	y	x^2	x^3	x^4	$x \cdot y$	$x^2 \cdot y$
1	1	1	1	1	1	1
2	4	4	8	16	8	16
3	9	9	27	81	27	81
4	15	16	64	256	60	240
Sum	**29**	**30**	**100**	**354**	**96**	**338**

Now, let's calculate each of the necessary sums from the table:

- $\sum y = 1 + 4 + 9 + 15 = 29$
- $\sum x = 1 + 2 + 3 + 4 = 10$
- $\sum x^2 = 1 + 4 + 9 + 16 = 30$
- $\sum x^3 = 1 + 8 + 27 + 64 = 100$
- $\sum x^4 = 1 + 16 + 81 + 256 = 354$
- $\sum x \cdot y = 1 + 8 + 27 + 60 = 96$
- $\sum x^2 \cdot y = 1 + 16 + 81 + 240 = 338$

Substitute the sums into equations:

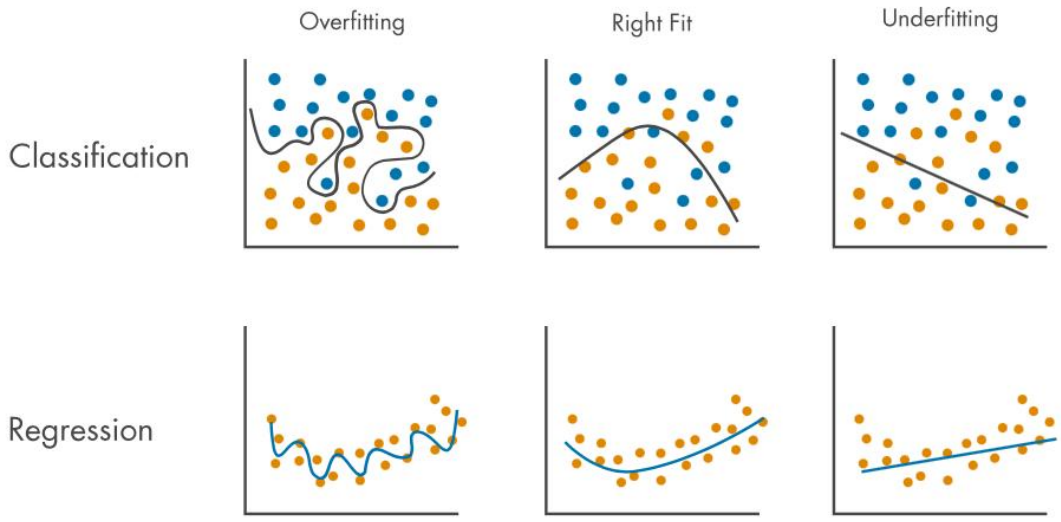
$$29 = 4a_0 + 10a_1 + 30a_2$$
$$96 = 10a_0 + 30a_1 + 100a_2$$
$$338 = 30a_0 + 100a_1 + 354a_2$$

Solving these we get,a_0=−0.75, a_1=0.95 a_2=0.75.
Thus, the quardatic equation becomes $y=-0.75+0.95x+0.75x^2$

For the given polynomial $y = -0.75 + 0.95x + 0.75x^2$, the corresponding y-values for $x = 1, 2, 3, 4, 5$ are:

- $x = 1, y = 0.95$
- $x = 2, y = 4.15$
- $x = 3, y = 8.85$
- $x = 4, y = 15.05$
- $x = 5, y = 22.75$

Overfitting: Overfitting occurs when a model is too complex and captures not only the underlying data patterns but also the noise or random fluctuations. It results in very high accuracy on the training data but poor generalization to new, unseen data.
Underfitting: Underfitting occurs when a model is too simple to capture the underlying patterns in the data. It results in poor performance on both the training and test data because the model makes oversimplified assumptions.



Avoiding overfitting in polynomial regression:

- ✓ Choose an appropriate degree for the polynomial: Higher-degree polynomials may fit the training data perfectly, but they can also introduce excessive noise, resulting in overfitting. Start with a lower-degree polynomial and gradually increase the degree while monitoring the performance on both the training and validation datasets.

- ✓ Regularization: Regularization methods help combat overfitting by adding penalty terms to the loss function. Two common techniques are L2 (Ridge) Regularization and L1 (Lasso) Regularization

- ✓ Cross-validation: Use k-fold cross-validation to evaluate the model's performance on multiple subsets of the data. This allows you to detect if your model is overfitting by checking if the model performs poorly on validation sets while performing well on the training set.

- ✓ Feature scaling: Polynomial regression can become more prone to overfitting if the input features have very different scales. Standardize or

normalize the input features to ensure that the higher-order terms do not dominate the model due to their scale.

- ✓ Early stopping: If you are using an iterative optimization method (like gradient descent), you can use early stopping to prevent the model from overfitting. Monitor the validation loss, and stop training when the validation performance stops improving.

- ✓ Increase the amount of data: A common cause of overfitting is insufficient data. The more data you have, the less likely the model will memorize the training data. You can collect more data or generate synthetic data to improve model generalization.

Categorical Regressors and Indicator Variables

Especially in regression, categorical regressors and indicator (dummy) variables are crucial for including categorical data (i.e., non-numeric data such as gender, region, or product type) in a model.

Categorical Regressors:

A categorical regressor (also called a categorical variable or factor) is a variable that represents categories or groups.

These categories might be:

- ✓ Nominal (no inherent order, e.g., colors like "Red", "Blue", "Green")
- ✓ Ordinal (with inherent order, e.g., "Low", "Medium", "High")

For instance, a variable such as "Region" might have categories like "North", "South", "East", and "West".

Regression models, such as linear regression, typically expect numeric input. Therefore, categorical variables need to be transformed into a numerical format to be included in these models.

Indicator (Dummy) Variables:

To represent categorical variables numerically, indicator variables (or dummy variables) are created.

Indicator variables (or dummy variables) are binary (0 or 1) variables created to represent categorical variables. The process of converting a categorical variable into dummy variables is known as dummy encoding.

How to Create Indicator Variables

1. One-hot Encoding: Each category of the categorical variable gets its own binary column. For example, for a categorical variable Color with three categories (Red, Blue, Green), we create three indicator variables:

- ✓ Color_Red: 1 if the color is red, 0 otherwise.
- ✓ Color_Blue: 1 if the color is blue, 0 otherwise.
- ✓ Color_Green: 1 if the color is green, 0 otherwise.

Example:

Original Data

Product	Color
Product1	Red
Product2	Blue
Product3	Green
Product4	Red

After One-Hot Encoding

Product	Color_Red	Color_Blue	Color_Green
Product1	1	0	0
Product2	0	1	0
Product3	0	0	1
Product4	1	0	0

2. Avoiding the Dummy Variable Trap:

To avoid perfect multicollinearity (where the variables are perfectly correlated), we usually exclude one of the dummy variables and treat it as the reference category. This category is implicitly captured when all dummy variables are 0. For the Color variable above, we might omit Color_Red, so the remaining dummy variables are: Color_Blue, Color_Green

Original Data

Product	Color
Product1	Red
Product2	Blue
Product3	Green
Product4	Red

After, avoiding dummy trap,

Product	Color_Blue	Color_Green
Product1	0	0
Product2	1	0
Product3	0	1
Product4	0	0

Note: When using a categorical variable with k levels (categories), you create k −1 dummy variables to avoid perfect multicollinearity.

Selection of variables and model building:

Variable selection means choosing among many variables which to include in a particular model, that is, to select appropriate variables from a complete list of variables by removing those that are irrelevant or redundant. The purpose of such selection is to determine a set of variables that will provide the best fit for the model so that accurate predictions can be made.

While choosing variables, there are two possible options:
1. In order to make the model as realistic as possible, the analyst may include as many as possible variables.
2. In order to make the model as simple as possible, one way includes only a fewer number of variables.

Both approaches have their consequences. In fact, model building and variable selection have contradicting objectives. When a large number of variables are included in the model, then these factors can influence the prediction of the study variable y . On the other hand, when a small number of variables are included then the predictive variance of $\hat{y}$ decreases.

Types of Variable Selection Techniques

- ✓ Filter Methods: (Select variables before model training)Filter methods apply a statistical measure to allocate a value to each variable . The variable are ranked based on those measure and also selected or removed from the dataset. (independent of the model).
  Examples:
  1. Correlation Analysis
  2. Chi-Square Test
  3. Mutual Information
  4. Variance Threshold

  These methods are simple, computationally efficient. But they may miss important interactions between variables in model.

- ✓ Wrapper Methods: Select features based on model performance.
  It includes:
  1. Forward selection
  2. Backward elimination
  3.Stepwise selection

  They have interactions between variables in model. But are computationally expensive.

✓ Embedded Methods: Embedded methods study which features best donate to the accuracy of the model i.e Feature selection occurs during model training.

1. LASSO (L1 Regularization)
2. Decision Trees/Random Forest feature importance
3. Elastic Net
4. Ridge Regression

These methods balances performance and computational cost.

✓ Some other methods includes Akaike Information Criterion(AIC) and Bayesian Information Criterion (BIC).

Feature selection techniques

1. Forward selection:

It is a stepwise approach used to build a regression model by starting with no predictors (variables) in the model, and then adding them one by one, based on their statistical significance.

- Start with an empty model: Initially, no variables are included in the model.
- Evaluate the potential variables: At each step, each variable not yet included in the model is tested for inclusion. The test is usually based on its p-value or test statistic.
- Select the most significant variable: The variable with the highest test statistic is chosen, provided it meets the predefined significance threshold.
- Update the model: The selected variable is added to the model, and the model is refitted.
- Recompute significance: After adding a variable, the significance of the remaining variables is recalculated.
- Repeat the process: The procedure continues until no remaining variable significantly contributes to the model at the cutoff level.
- Final model: The model consists of only the variables that have a significant contribution, and once a variable is added, it stays in the model throughout.

2. Backward elimination: Backward elimination is the simplest of all variable selection methods.

- Start with a full model: Include all variables initially.
- Test significance of variables: Evaluate the test statistics or p-values for each variable.

- Remove the least significant variable: Identify the variable with the smallest test statistic or the highest p-value above the cut-off value (e.g., greater than 0.05).
- Refit the model: Remove the identified variable and re-compute the test statistics or p-values for the remaining variables.
- Repeat the process: Continue deleting the least significant variable (based on the smallest test statistic or the highest p-value above the cut-off value).
- Stop when only significant variables remain: The process continues until all remaining variables are significant at the cut-off level.
- p-to-remove: The cut-off value associated with p-value doesn't have to be set to 0.05, and can be adjusted based on the model's needs.

This method ensures that only the most significant predictors remain in the model.

3. Stepwise selection:

It is a hybrid method that combines both forward and backward selection.

- Combination of forward and backward selection: Stepwise selection allows adding and removing variables in both directions.
- Starting with forward selection: The process can begin with forward selection, where variables are added one at a time based on statistical significance.
- After each addition, check for insignificant variables: Once a variable is added, the procedure checks all the variables already included in the model.
- Remove insignificant variables: If any of the variables in the model are no longer significant, they are removed.
- Repeat until no further changes: The process continues until all variables in the model are significant and all excluded variables are insignificant.
- Modified forward selection: Due to its nature, stepwise selection can be viewed as a modified forward selection, as it incorporates both addition and removal of variables during the process.

This method balances both adding and removing variables to find the most optimal model.

Model Building:

Once relevant features are selected, you proceed to train a predictive model.

According to Hafermann et al. (2021) and Greenland and Pearce (2015), there are three main strategies for building models:

1. Adjust All: Include all potential confounding variables in the model.
A confounding variable is something that can distort or confuse the relationship you're studying, so including these variables helps control for their effects.
2. Predictor Selection: Choose variables based on how well they predict the outcome or exposure (or both) in the model.
3. Change in Estimate (CIE) Selection: Involves deciding which variables to include in your model based on their impact on the estimate of the exposure effect. Begin by including only the key variables (like the exposure of interest and any "forced" variables such as age or sex). Add one variable at a time to the model and see how much it changes the estimate of the effect of the exposure on the outcome. If the change is very small, the variable is excluded because it doesn't meaningfully affect the results.

Also, Before building a model, it's important to complete the following steps:

- ✓ Thoroughly check, describe, and summarize the data.
- ✓ Adjust quantitative variables:
  - Rescale variables: For eg. If weight is measured in grams, rescale it to kilograms to make the numbers more interpretable.
  - Re-center variables: Shift numerical variables so that a value of zero is meaningful within the dataset.Imagine you're studying poverty and the effect of income on well-being. In your dataset, income is recorded in dollars, but the income range spans from $0 to $1,000,000.
  However, zero income doesn't necessarily represent a meaningful reference point in your study. Instead, it might make more sense to re-center income by subtracting the median income of the population (say, $50,000) from all income values.
  Now, the new "zero point" is the median income, meaning any value above or below zero indicates how much higher or lower someone's income is compared to the median.
- ✓ Use contextual knowledge to select categories or flexible forms (like splines) for detailed modeling.
  - Contextual knowledge: refers to understanding the subject you're studying and using that knowledge to make decisions about how to handle the data.
  For example, if you're studying age, you might know from your subject

knowledge that people often group themselves in age ranges. So instead of using exact ages like 22, 35, or 67 in your model, you might group them into age categorieslike: "10-19 years", "20-29 years", "50+ years". This makes the data easier to interpret and analyze.

- Flexible techniques like splines:

Sometimes, the relationship between two variables is not linear (i.e., it doesn't follow a straight line). For example, the relationship between age and income might not increase smoothly as age goes up. Instead, it might go up quickly in early career years, then level off or drop later in life.

Splines are mathematical tools that let you create smooth, curved lines rather than forcing everything into a straight-line relationship. This way, you can capture more complex patterns.

Regression Metrics: MSE, RMSE, MAE, R^2, Adjusted R^2

Different metrics can be used to evaluate the performance of regression models, and the choice of metric depends on the specific problem and goals of the analysis. Here are some common regression evaluation metrics and their applications:

1. Mean Squared Error (MSE): This is a commonly used metric that measures the average of the squared differences between the predicted and actual values. It gives more weight to large errors and is sensitive to outliers. Lower MSE indicates better model performance.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where, MSE = Mean Square error, n = Number of Data points, Y_i = Observed Values, $\hat{Y}_i$ = Predicted Values

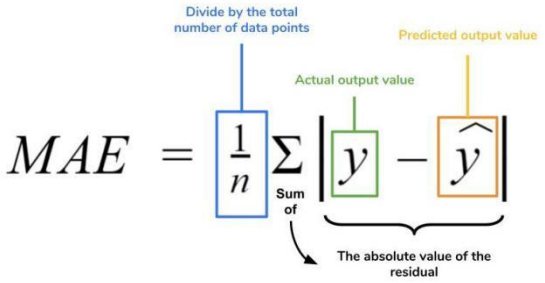
2. Root Mean Squared Error (RMSE): This metric is the square root of MSE and has the same interpretation. It is easier to interpret since the units of measurement match the target variable. RMSE is useful when the goal is to minimize the overall error in the model. Like MSE, lower RMSE indicates a better fit and easier to interpret than MSE as it matches the scale of y.

$$\text{RMSE} = \sqrt{MSE}$$

where

MSE = mean squared error

3. Mean Absolute Error (MAE): This metric measures the average of the absolute differences between the predicted and actual values. It gives equal weight to all errors and is less sensitive to outliers. MAE is useful when the goal is to minimize the overall error in the model while avoiding large errors. It is a more robust metric than MSE or RMSE for datasets with outliers.



$$MAE = \frac{1}{n}\sum \left| y - \hat{y} \right|$$

4. R-squared (R^2): This metric measures the proportion of variance in the target variable explained by the model. It ranges from 0 to 1, with higher values indicating better performance. R^2 is useful when the goal is to explain the variability in the target variable using the predictors.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

5. Adjusted R-squared (Adj R^2): This metric is similar to R^2 but accounts for the number of predictors used in the model. It penalizes overfitting by adjusting for the number of predictors. Adj R^2 is useful when the goal is to explain the variability in the target variable using a parsimonious model(a model that explains variability in the target variable with as few predictors as possible.)

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where

R^2 Sample R-Squared

N Total Sample Size

p Number of independent variable

Model selection metrics:

1. AIC (Akaike Information Criterion)

   The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from:

$$AIC = -2\ln(L) + 2k$$

where, k: Number of parameters in the model and L: Maximized value of the likelihood function for the model.
Note: Lower AIC values indicate a better model.

2. BIC (Bayesian Information Criterion)

Similar to AIC, the Bayesian Information Criterion (BIC) is another model selection criterion that considers both model fit and complexity. BIC is based on Bayesian principles and provides a more stronger penalty for model complexity compared to AIC.

$$\text{BIC} = k \ln(n) - 2\ln(L)$$

Where:

- k: Number of parameters in the model.
- n: Sample size.
- L: Maximized value of the likelihood function for the model.

Note: Lower BIC values indicate a better model.

Example Dataset:

X1	X2	Y(Actual)	$\hat{Y}$ (Predicted)
1	2	6	5.8
2	3	8	7.9
3	4	10	9.4
4	5	12	11.1
5	6	14	13.3

1. Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Substitute the values:

$$\text{MSE} = \frac{1}{5}\left[(6-5.8)^2 + (8-7.9)^2 + (10-9.4)^2 + (12-11.1)^2 + (14-13.3)^2\right]$$

$$\text{MSE} = \frac{1}{5}\left[0.04 + 0.01 + 0.36 + 0.01 + 0.49\right]$$

$$\text{MSE} = \frac{0.91}{5} = 0.182$$

2. Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{0.182} \approx 0.426$$

3. Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Substitute the values:

$$\text{MAE} = \frac{1}{5}\left[|6-5.8| + |8-7.9| + |10-9.4| + |12-11.1| + |14-13.3|\right]$$

$$\text{MAE} = \frac{1}{5}\left[0.2 + 0.1 + 0.6 + 0.9 + 0.7\right]$$

$$\text{MAE} = \frac{2.5}{5} = 0.5$$

4. R-squared (R^2)

First, calculate the mean of the actual values Y:

$$\bar{y} = \frac{6+8+10+12+14}{5} = 10$$

Now calculate the sum of squared residuals (SSR) and total sum of squares (SST):

$$\text{SSR} = (6-5.8)^2 + (8-7.9)^2 + (10-9.4)^2 + (12-11.1)^2 + (14-13.3)^2$$

$$\text{SST} = (6-10)^2 + (8-10)^2 + (10-10)^2 + (12-10)^2 + (14-10)^2$$

Now calculate R^2:

$$R^2 = 1 - \frac{0.91}{40} = 1 - 0.02275 = 0.97725$$

5. Adjusted R-squared (Adjusted R^2)

The formula for **Adjusted R^2** is:

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-P-1}\right)$$

Substitute the values:

$$1 - R^2 = 1 - 0.97725 = 0.02275$$

$$n - P - 1 = 5 - 2 - 1 = 2$$

Now calculate Adjusted R^2:

$$\text{Adjusted } R^2 = 1 - \left(\frac{(0.02275)(4)}{2}\right) = 1 - \left(\frac{0.091}{2}\right) = 1 - 0.0455 = 0.9545$$

6. Akaike Information Criterion (AIC)

For this calculation, we'll assume the log-likelihood L is approximately $-n \times \text{MSE}/2$:

$$\ln(L) \approx -5 \times 0.182/2 = -0.455$$

Now calculate AIC:

$$AIC = 2P - 2\ln(L) = 2(2) - 2(-0.455) = 4 + 0.91 = 4.91$$

7. Bayesian Information Criterion (BIC)

$$BIC = \ln(n)P - 2\ln(L)$$

Substitute the values:

$$BIC = \ln(5)(2) - 2(-0.455) = 1.609 \times 2 + 0.91 = 3.218 + 0.91 = 4.128$$

Assignment :

Consider the following dataset where X1, X2, and X3are three independent variables, and Yis the dependent variable:

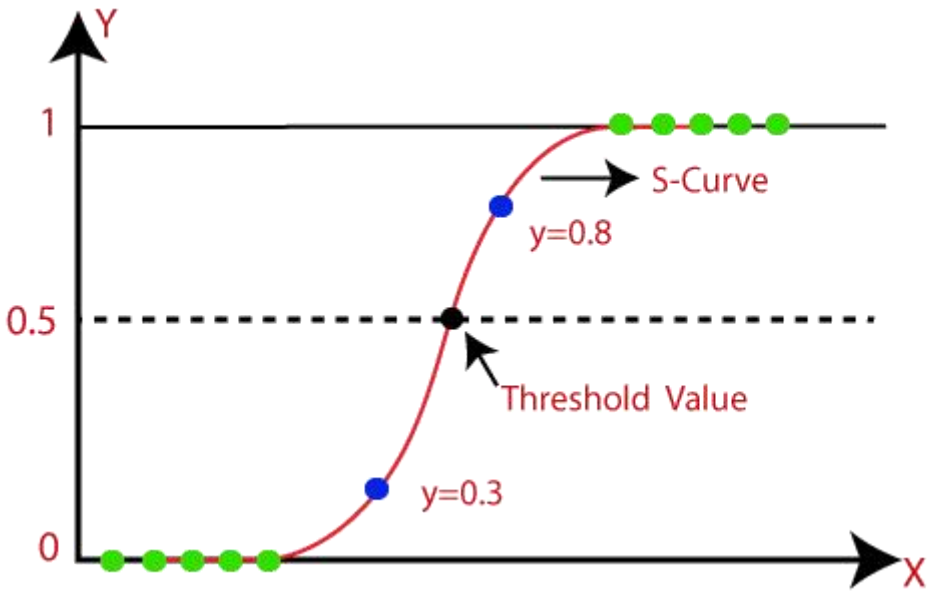
X1	X2	X3	Y(Actual)	$\hat{Y}$ (Predicted)
1	2	2	10	9.5
4	3	3	12	11.8
2	5	5	14	13.6
6	1	1	16	15.4
3	4	4	8	8.2

Calculate Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), Adjusted R-squared (Adjusted R^2), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)

Logistic Regression

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.



We all know the equation of the best fit line in linear regression is:

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

Let's say instead of y we are taking probabilities (P).

$$P = \beta_0 + \beta_1 x$$

But there is an issue here, the value of (P) will exceed 1 or go below 0 and we know that range of Probability is (0-1). To overcome this issue we take "odds" of P:

Odds are a way of expressing probabilities. They represent the ratio of the probability of success (p) to the probability of failure (1−p).

$$\text{Odds} = \frac{p}{1-p}$$

For example:

- If the probability of success (p) is 0.75, the odds are:

$$\text{Odds} = \frac{0.75}{1 - 0.75} = 3$$

This means the event is 3 times more likely to occur than not occur.

So,

$$\frac{P}{1-P} = \beta_0 + \beta_1 x$$

Also, Odds are always positive $(0,+\infty)$, but this restricted range complicates modeling i.e it is difficult to model a variable that has a restricted range. To control this we take the log of odds which has a range from $(-\infty,+\infty)$.

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$$

Now we just want a function of P because we want to predict probability right? not log of odds. To do so we will multiply by exponent on both sides and then solve for P.

$$\exp[\log(\frac{p}{1-p})] = \exp(\beta_0 + \beta_1 x)$$

$$e^{\ln[\frac{p}{1-p}]} = e^{(\beta_0 + \beta_1 x)}$$

$$\frac{p}{1-p} = e^{\left(\beta_0 + \beta_1 x\right)}$$

$$p = e^{\left(\beta_0 + \beta_1 x\right)} - pe^{\left(\beta_0 + \beta_1 x\right)}$$

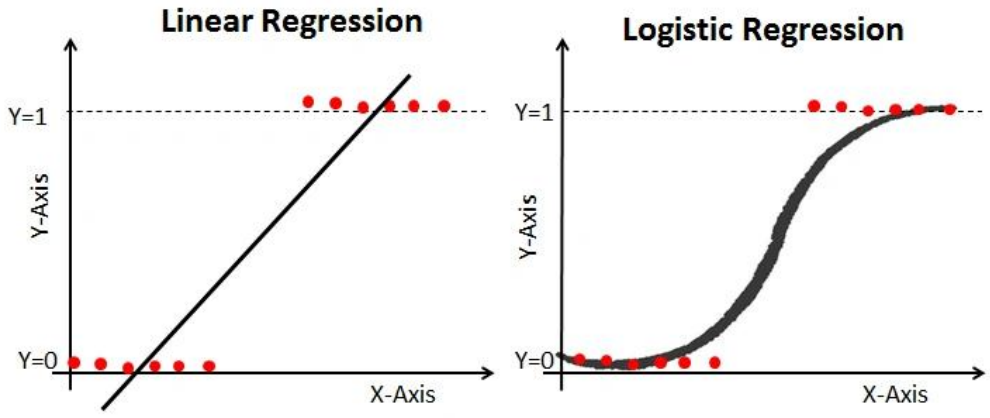
$$p[1 + e^{\left(\beta_0 + \beta_1 x\right)}] = e^{\left(\beta_0 + \beta_1 x\right)}$$

$$p = \frac{e^{\left(\beta_0 + \beta_1 x\right)}}{1 + e^{\left(\beta_0 + \beta_1 x\right)}}$$

$$Now\ dividing\ by\ e^{(\beta_0+\beta_1 x)},\ we\ will\ get$$

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}\ This\ is\ our\ sigmoid\ function.$$

Now we have our logistic function, also called a sigmoid function. The graph of a sigmoid function is as shown below. It squeezes a straight line into an S-curve.



Self Study: Differences Between Linear and Logistic Regression

For more info: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/

$if\ p < 0.5\ then\ \text{N}\ else\ \text{Y}$

BUSAGE	DAYSDELQ	DEFAULT	y'	p	Prediction
87	2	N	-4.811	0.008	N
89	2	N	-4.795	0.008	N
100	26	Y	-2.261	0.094	N
275	54	Y	1.983	0.879	Y
42	57	Y	0.437	0.608	Y
88	53	N	0.395	0.597	Y

$$y' = -5.706 + 0.008 \times 87 + 0.102 \times 2 = -4.811$$

$$p = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2)}} = \frac{1}{1+e^{-y'}}$$

$$p = \frac{1}{1+e^{4.811}} = 0.008 \xrightarrow{<0.5} \text{N}$$

Example: The dataset of pass or fail in an exam of 5 students in the given table:

Hours Study	Pass(1)/ Fail(0)
29	0
15	0
33	1
28	1
39	1

Use logistic regression to answer the following questions.
1. Calculate the probability of pass for the student who studied 33 hours.
2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%
Assume Odds of passing the course is log(odds)= -64 +2*hours

Solution,

1: Probability for a student who studied 33 hours

The relationship is given as:

$$\log(\text{odds}) = -64 + 2 \times \text{hours}.$$

For $\text{hours} = 33$:

$$\log(\text{odds}) = -64 + 2 \times 33 = -64 + 66 = 2.$$

The odds are:

$$\text{odds} = e^{\log(\text{odds})} = e^2 \approx 7.389.$$

The probability is:

$$p = \frac{\text{odds}}{1 + \text{odds}} = \frac{7.389}{1 + 7.389} \approx \frac{7.389}{8.389} \approx 0.881.$$

So, the probability of passing for a student who studied 33 hours is approximately **88.1%**.

2: Minimum hours required for $p > 0.95$

To find the minimum hours for a probability of 0.95, we start with the formula for the odds:

$$\text{odds} = \frac{p}{1-p}.$$

For $p = 0.95$:

$$\text{odds} = \frac{0.95}{1-0.95} = \frac{0.95}{0.05} = 19.$$

Now, take the logarithm of the odds:

$$\log(\text{odds}) = \log(19) \approx 2.944.$$

The relationship between log(odds) and hours is:

$$\log(\text{odds}) = -64 + 2 \times \text{hours}.$$

Substitute $\log(\text{odds}) = 2.944$:

$$2.944 = -64 + 2 \times \text{hours}.$$

Solve for hours:

$$2 \times \text{hours} = 2.944 + 64 = 66.944,$$

$$\text{hours} = \frac{66.944}{2} \approx 33.472.$$

So, a student must study at least **33.5 hours** to have a probability of passing greater than 95%.

Case Studies:

1. Healthcare: Predicting Disease Outcomes

A case study explored the use of logistic regression to Predict whether a patient has diabetes based on factors such as age, BMI, blood pressure, and glucose levels.

- Dataset: Pima Indians Diabetes Dataset.
- Features: Age, BMI, number of pregnancies, glucose concentration, etc.
- Outcome Variable: Binary (1 = Diabetic, 0 = Non-Diabetic).

- Model: Logistic regression was applied to estimate the probability of diabetes.
- Results: Accuracy, precision, recall, and ROC-AUC scores were used to evaluate the model's performance.

It helps us to identify high-risk patients early, enabling timely intervention and resource allocation.

2. Marketing: Customer Churn Prediction

A marketing case study uses Logistic Regression to identify customers likely to cancel their subscriptions.

- Dataset: Customer behavior data from a telecom company.
- Features: Monthly charges, contract type, tenure, support requests, etc.
- Outcome Variable: Binary (1 = Churn, 0 = Retain).
- Model: Logistic regression was used due to its interpretability and ease of implementation.
- Evaluation: Metrics such as F1-score and ROC-AUC were prioritized.

This helps to increase customer retention through targeted campaigns for at-risk customers.

3. Education: Predicting Student Performance

In education, it is used to predict whether a student will pass or fail a course.

- Dataset: Academic performance data including attendance, assignment scores, and participation in class activities.
- Outcome Variable: Binary (1 = Pass, 0 = Fail).
- Model: Logistic regression to identify key factors affecting student performance.

It helps educators provide targeted support to students at risk of failing.

4. Human Resources:

It can be used to predict whether an employee will leave the company.

- Dataset: HR data containing features like job satisfaction, salary, working hours, and years at the company.
- Outcome Variable: Binary (1 = Leave, 0 = Stay).
- Model: Logistic regression provided interpretable coefficients for HR teams to understand the driving factors of attrition.

It enables companies to implement retention strategies for employees at risk of leaving.