We acknowledge the Australian Aboriginal and Torres Strait Islander peoples as the traditional owners of the lands and waters where we live and work.

Week 1: Introduction to Data Mining

Dr Max Cao
Performance and Insights Manager at Brisbane City Council & Lecturer at James Cook University

Subject Assessments

Quiz 1: Online quiz (Non-credit)

Due Sunday Week 1 (19th Jan 2025)

Assessment 1: Feature selection with Naïve Bayes, Discriminant Analysis (30%)

Due Sunday Week 3 (2nd Feb 2025)

Assessment 2: Classification and Clustering (30%)

Due Sunday Week 5 (16th Feb 2025)

Assessment 3: Capstone project (40%)

Due Wednesday Week 7 (26th Feb 2025)

To pass this subject, you must:

Achieve an overall percentage of 50% or more.

Achieve a percentage of 50% or more in the capstone project

Subject content

Week 1: Introduction, Linear Regression and R markdown.

Week 2: Supervised Learning – Bayes Classifiers.

Week 3: Supervised Learning – kNN and Logistic Regression.

Week 4: Unsupervised Learning – Clustering.

Week 5: Unsupervised Learning – Outlier Detection and PCA.

Week 6: Notions of Basket Analysis and Recommender Systems.

Data Mining

The process of discovering useful patterns, information, knowledge and predictive models from larger-scale data by

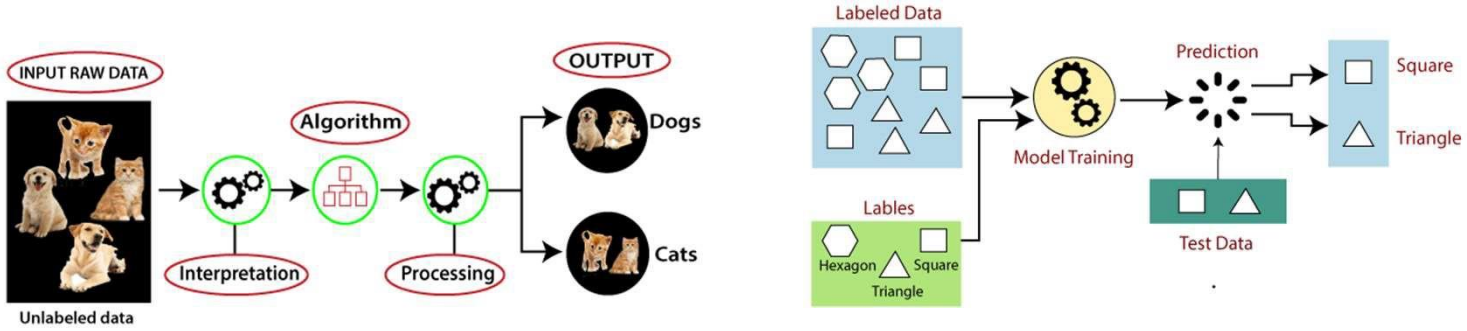
- Applications of the machine learning techniques
- Implementation of the techniques using R
- Statistical theory/assumptions underlying the techniques

Data Mining categorization

Supervised learning	Unsupervised learning
A set of p features $X_1, X_2, \ldots, X_p$ measured on n observations	A set of p features $X_1, X_2, \ldots, X_p$ measured on n observations.
and a response Y measure on those same n observations	There is no associated response variable Y
The goal is to predict Y using $X_1, \ldots, X_p$	The goal is to discover interesting things about the measurements on $X_1, X_2, \ldots, X_p$

Dataset Identifications

Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision, whereas supervised learning has labelled dataset.



Source https://www.javatpoint.com

Supervised Learning

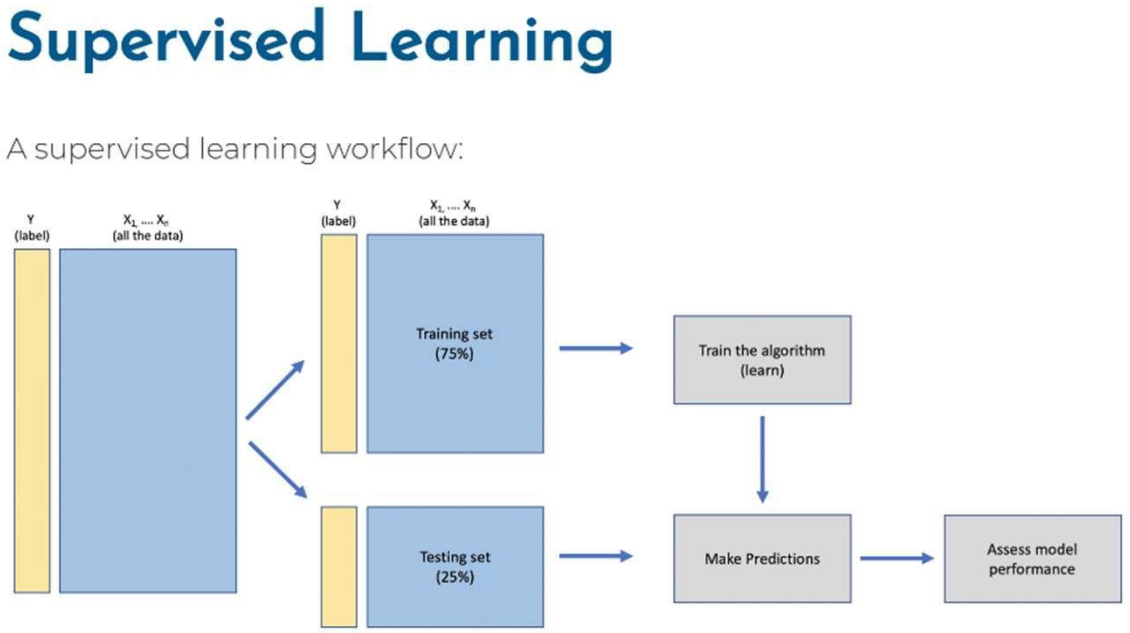
Look out factors-Details in slide 21 and 22

Bias- Variance Trade-off

Train and Test split

Confusion Matrix

Overfitting/Underfitting? Why?



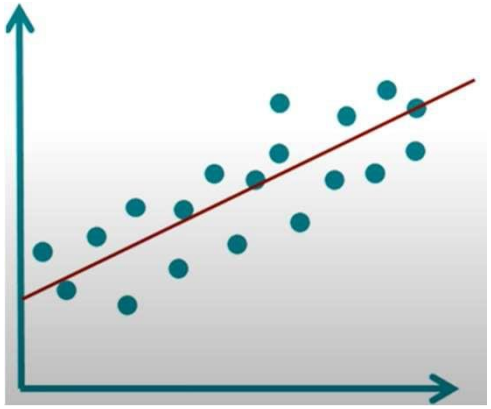
Data Mining Goals

There are numerous problems that data scientists may aim to solve in the realm of data mining, but the majority can be classified into six core tasks/goals.

- Regression.
- Classification.
- Anomaly/outlier detection.
- Clustering.
- Association rules.

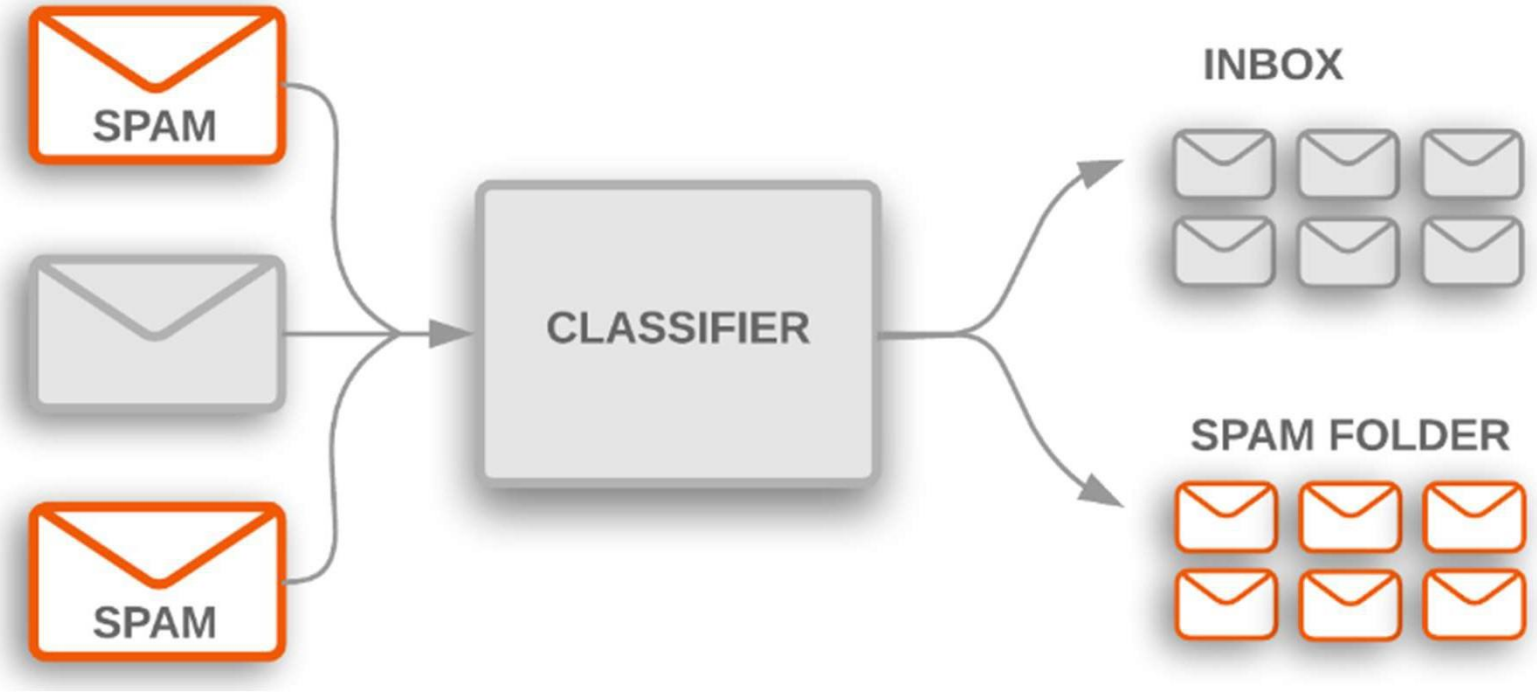
Regression

Regression is used to explore the relationship between a dependent variable (denoted as Y) and one or number of explanatory variables (denoted as X), represented as a straight line.



Classification

Classification is the problem of identifying to which of a set of categories an observation belongs to.



Aassociation Rules

Association, is an analytical process that finds frequent patterns, associations, or causal structures from data.

It is an approach to data, and as the basis for finding association rules. It can also be used to find discriminative features for classification or clustering. How frequently milk and bread go together in a dataset, is a frequent associated itemset/sub sequences /substructures.

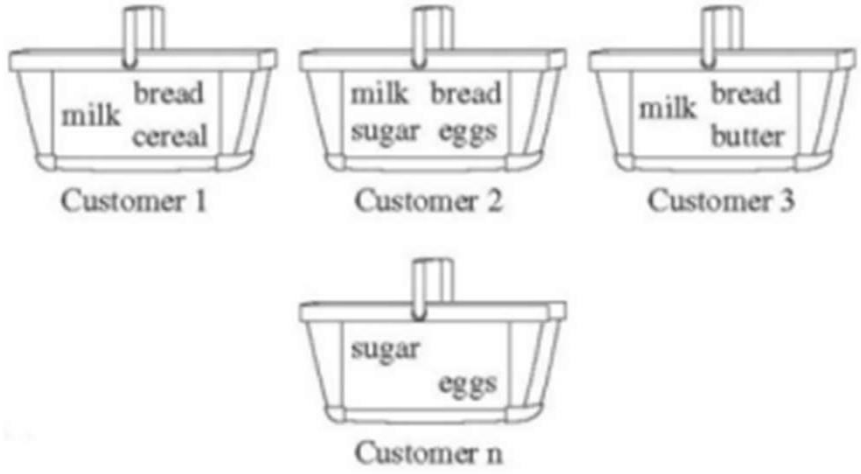
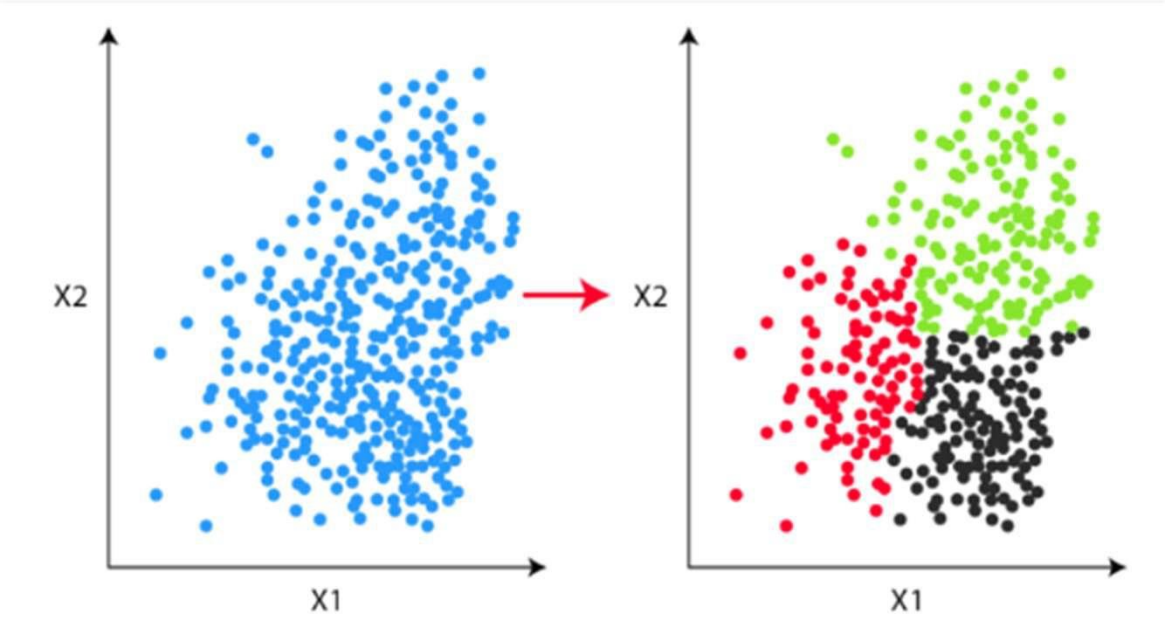


Fig.1: Market Basket Analysis (Han, Kamber & Pei).

Clustering

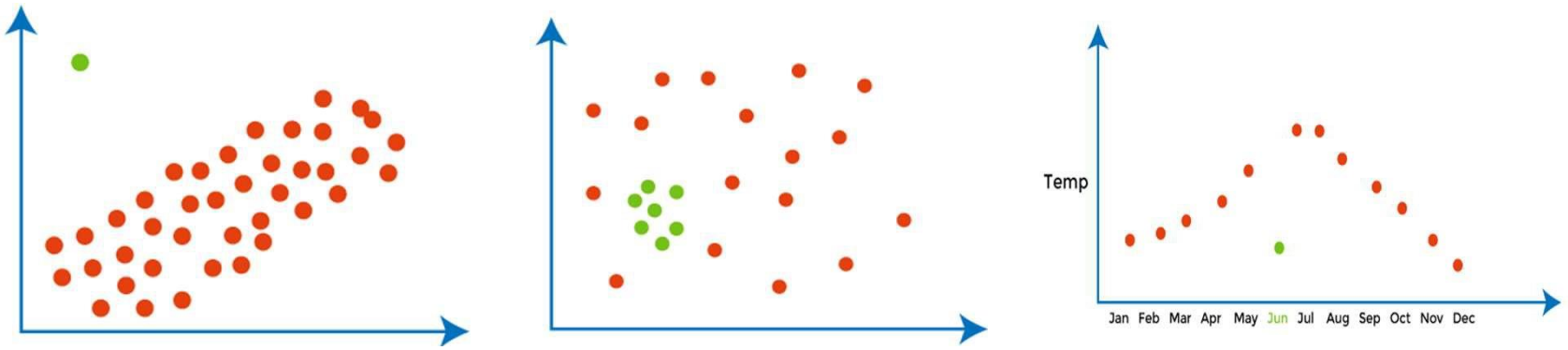
Clustering is used to find groups (clusters) of observations that are more similar or related to each other than observations in other groups.



Anomaly/outlier detection

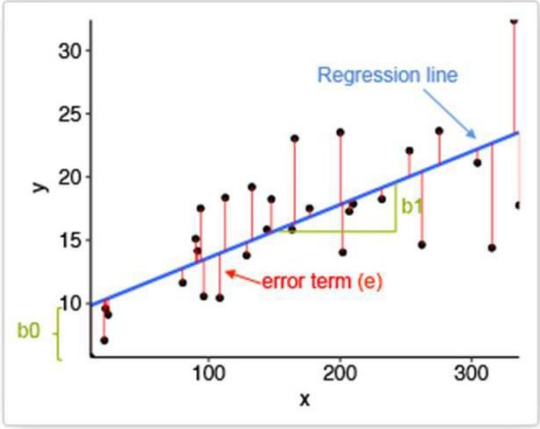
Anomaly/outlier detection is used to discriminate between ‘normal’ and ‘anomalous’ observations.

Outliers can have a significant impact on the accuracy of machine learning models. If outliers are not detected and handled appropriately, they can lead to overfitting, underfitting, or biased results. For instance, in fraud detection, failing to detect fraudulent transactions can lead to significant financial losses. In the medical field, failing to detect an outlier in patient data can result in incorrect diagnoses and treatments. Methods -Z-Score Method (no. of standard deviations a data point is away from the mean. We consider data points that have a Z-score greater than a certain threshold as outliers), Interquartile Range (IQR) Method, Local outlier Factor method(It identifies data points with significantly lower density than their neighbors as outliers).



Supervised learning-Regression

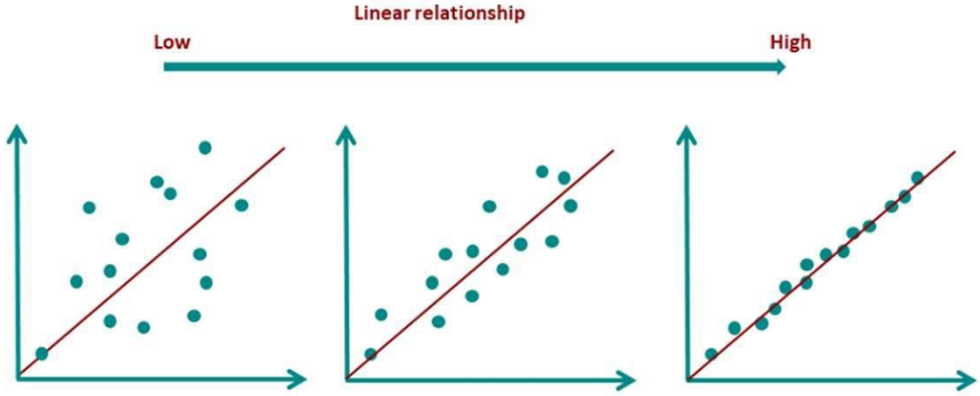
In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more explanatory variables. It is represented by **Y = a + bX + E**, where a = intercept, the value of y when x = 0, b = slope, X=explanatory variable and E is the error term(difference between true value and the estimated value). X can be both quantitative and qualitative. Y is continuous. The relationship between the two variables is approximated by a straight line. Overall, the residual errors (e) have approximately mean zero.



Regression Analysis goals

In order to predict the value of the dependent variable for individuals for whom some information concerning the explanatory variables is available(classification), or in order to estimate the effect of some explanatory variable on the dependent variable(Regression).

The greater the linear relationship between the dependent and independent variables, the more the data points lie on a straight line.

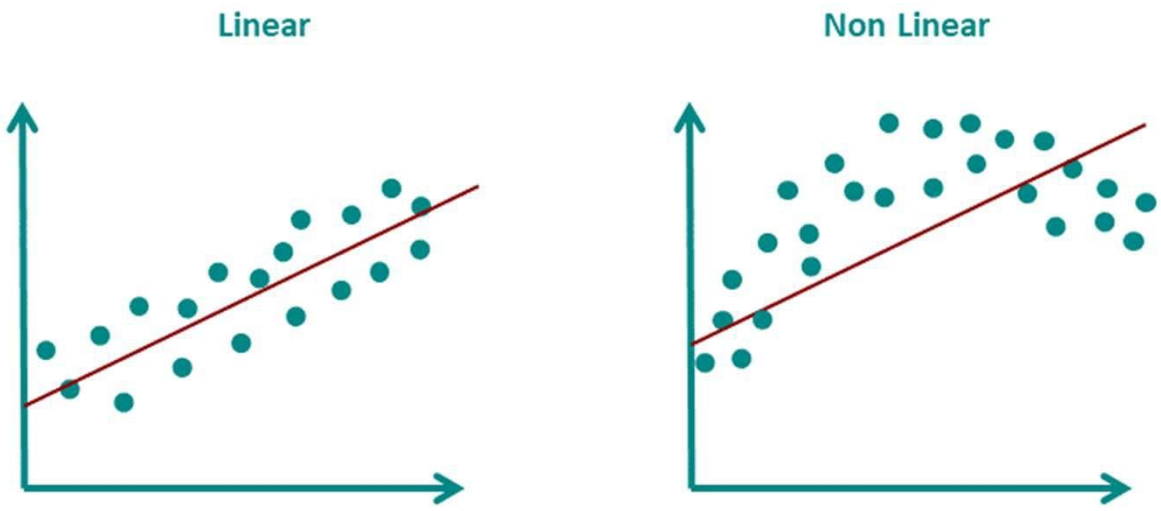


Assumptions- Regression

- **Linearity:** There must be a linear relationship between the dependent and independent variables.
- **Homoscedasticity:** The residuals must have a constant variance.
- **Normality:** Normally distributed error. Zero conditional mean $E(\varepsilon i) = 0$
- **No Multicollinearity:** No high correlation between the independent variables

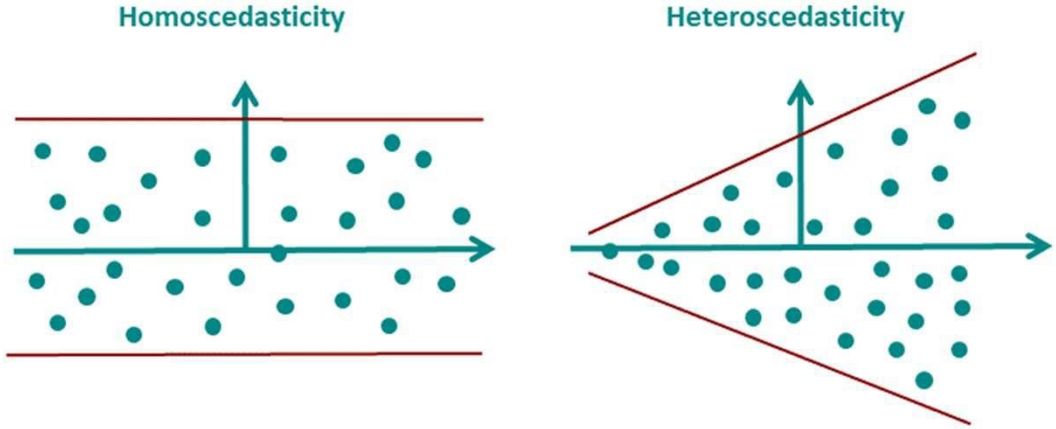
Linear

In linear regression, a straight line is drawn through the data. This straight line should represent all points as good as possible. If the points are distributed in a non-linear way, the straight line cannot fulfill this task.



Homoscedasticity

Since in practice the regression model never exactly predicts the dependen t variable, there is always an error. This very error must have a constant variance over the predicted range.



To test homoscedasticity, i.e. the constant variance of the residuals, the dependent variable is plotted on the x-axis and the error on the y-axis. Now the error should scatter evenly over the entire range. If this is the case, homoscedasticity is present. In the case of heteroscedasticity, the error has different variances, depending on the value range of the dependent variable.

Normal distribution of the error

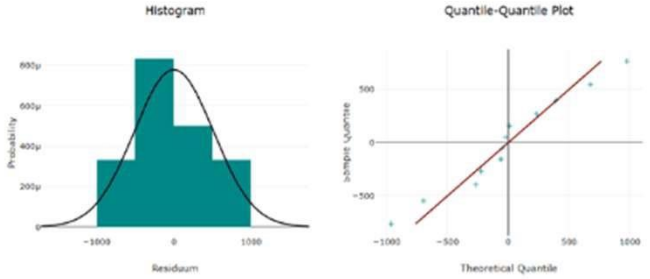
The next requirement of linear regression is that the error epsilon must be normally distributed. There are two ways to find it out: One is the analytical way and the other is the graphical way. In the analytical way, you can use either the Shapiro-Wilk test. If the p-value is greater than 0.05, there is no deviation of the data from the normal distribution and one can assume that the data are normally distributed.

Analytical

Copy

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistics	df	p-value	Statistics	df	p-value
Residuum	0,16	12	0,873	0,973	12	0,936

Graphically



Multicollinearity, means that two or more independent variables are strongly correlated with one another. The problem with multicollinearity is that the effects of each independent variable cannot be clearly separated from one another.

To check for multicollinearity problem, VIF stands for variance inflation factor. It measures how much the variance of any one of the coefficients is inflated due to multicollinearity in the overall model.

VIF (Variance Inflation Factor)

$$VIF = \frac{1}{1 - R^2}$$

R^2: Coefficient of determination

VIF Test

VIF equal to 1 = variables are not correlated.

VIF between 1 and 5 = variables are moderately correlated.

VIF greater than 5 = variables are highly correlated.

Predictions Errors-Bias-Variance Trade-off

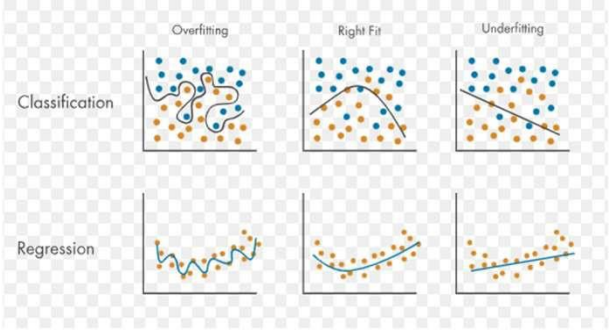
Bias is a systematic error that occurs due to wrong assumptions in the machine learning process.

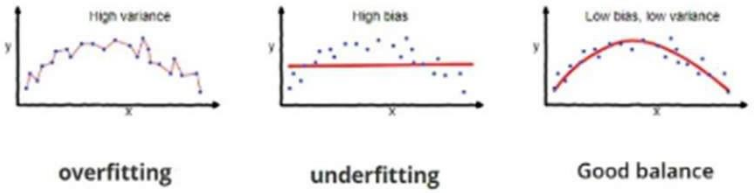
It is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data. The high-bias model will not be able to capture the dataset trend. It is considered as the **underfitting model,** which has a high error rate. It is due to a very simplified algorithm. An underfit model has poor performance on the training data and will result in unreliable predictions(new dataset). **For example,** a linear regression model may have a high bias if the data has a non-linear relationship.

Few ways to reduce -Increase the size of the training data, Increase the number of features, use a complex model.

Variance ,is a measure of how data points differ from the mean. More specifically, variance is the variability of the model that how much it is sensitive to another subset of the training dataset. i.e. how much it can adjust on the new subset of the training dataset(Overfitting).

Overfitting is an undesirable machine learning behavior that occurs when the machine learning model gives accurate predictions for training data but not for new data. **Few ways to reduce** -Cross validation(k-fold), Feature selection-relevant feature(variable importance), Simplifying the model/complexity. **"Bias and variance are complements of each other".** The increase of one will result in the decrease of the other and vice versa. Hence, finding the right balance of values is known as the **Bias-Variance Tradeoff**. An ideal algorithm should neither underfit nor overfit the data. Can we achieve a right fit model?





Simple Vs Multiple Linear Regression

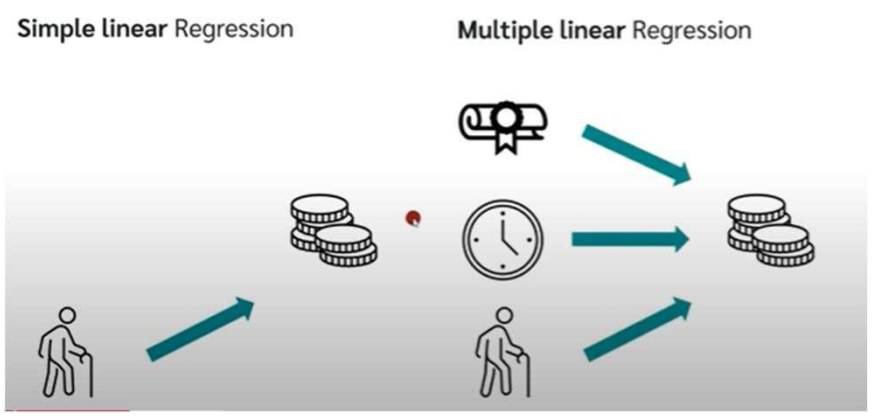
Simple Linear regression formula

$Y = a + bX + \varepsilon.$

Multiple Linear regression formula

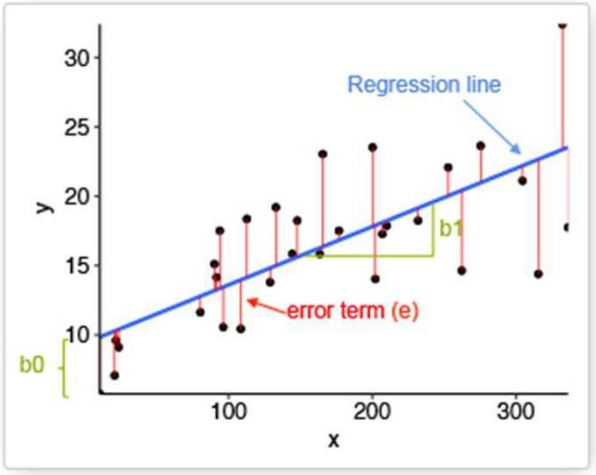
$Y = a + b_1X_1 + b_2X_{2....} b_nX_{n+} \boldsymbol{\varepsilon}$

Regression



Vs Classification

Weight	Height	Age	Gender
79	1.80	35	Male
69	1.68	39	Male
73	1.82	25	Male
95	1.70	60	Male
82	1.87	27	Male
55	1.55	18	Female
69	1.50	89	Female
71	1.78	42	Female
64	1.67	16	Female
69	1.64	52	Female



Straight Line/Line of best fit determination –Least Square

The sum of the squares of the residual errors are called the **Residual Sum of Squares** or **RSS**. The residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. The average variation of points around the fitted regression line is called the **Residual Standard Error** (**RSE**).

This is one the **metrics** used to evaluate the overall quality of the fitted regression model.

The lower the RSE, the better it is. Since the mean error term is zero, the outcome variable y can be approximately estimated as $y \sim b0 + b1*x$. The parameter β (the regression coefficient) signifies the amount by which change in x must be multiplied to give the corresponding average change in y, or the amount y changes for a unit increase in x.

Beta coefficients b0 and b1 are determined to understand how closely your data fit on a line, so that the RSS is as minimal as possible.

This method of determining the beta coefficients is technically called **least squares** regression or **ordinary least squares** (OLS) regression.

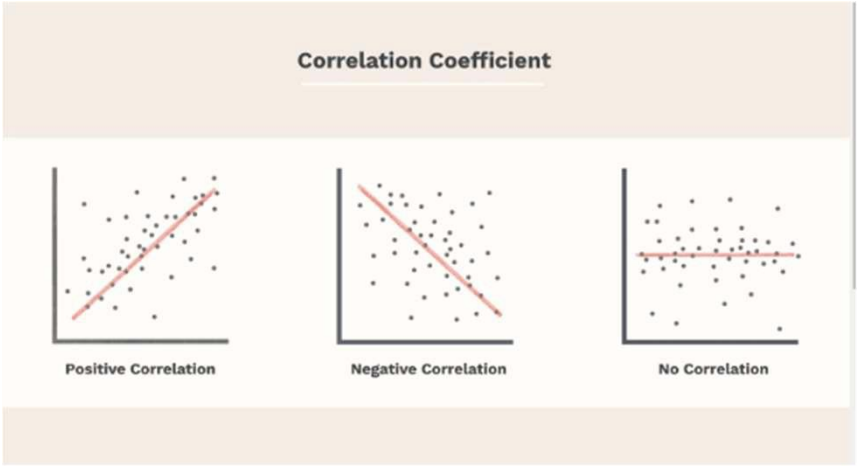
The **regression coefficient** *b* can now have different signs, ranges from -1 and +1.which can be interpreted as follows

- $b > 0$: there is a positive correlation between x and y (the greater x, the greater y)
- $b < 0$: there is a negative correlation between x and y (the greater x, the smaller y)
- $b = 0$: there is no correlation between x and y

Model Strength Analysis

A correlation or simple linear regression analysis can determine if two numeric variables are significantly linearly **related.** A correlation analysis provides information on the **strength** and **direction** of the linear relationship between two continuous variables. The classic one for regression analysis is **“Pearson Coefficient”.**

The Pearson correlation coefficient r_{XY} is a measure of the strength of the linear relationship between two variables X and Y and it takes values in the closed interval [−1, +1]. Describes the direction of the slope.



Multiple Linear Regression-Line of Best Fit

The coefficient of determination (R^2) measures how well a statistical model predicts an outcome. This is on the one hand the **coefficient of determination** R^2 and on the other hand the **standard estimation error.**

The coefficient of determination R^2, also known as the variance explanation, indicates how large the portion of the variance is that can be explained by the explanatory variables. The more variance can be explained, the better the regression model is. In order to calculate R^2, the variance of the estimated value is related to the variance in the observed values:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Where:

- RSS = sum of squared residuals
- TSS = total sum of squares

Coefficient of determination (R^2)	Interpretation
0	The model **does not** predict the outcome.
Between 0 and 1	The model **partially** predicts the outcome.
1	The model **perfectly** predicts the outcome.

Adjusted R2 –To fix the overestimation issue of the coefficient of determination when too independent variables are used. The coefficient of determination R^2 is influenced by the number of independent variables used. The more independent variables are included in the regression model, the greater the variance resolution R^2. To take this into account, the adjusted R^2 is used.

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \frac{MSE}{MST}$$

where p is the number of parameters

Standard estimation error: How much the model overestimates the dependent variable on an average.

Regression-Hypothesis Test

- **Null Hypothesis (H_0)** – This can be thought of as the implied hypothesis. "Null" meaning "nothing." This hypothesis states that there is no difference between groups or no relationship between variables.
- **Alternative Hypothesis (H_a)** – This is also known as the claim. This hypothesis should state what you expect the data to show, based on your research on the topic.

(Example)

Null Hypothesis: H_0: There is no relationship between height and shoe size.

Alternative Hypothesis: H_a: There is a positive relationship between height and shoe size.

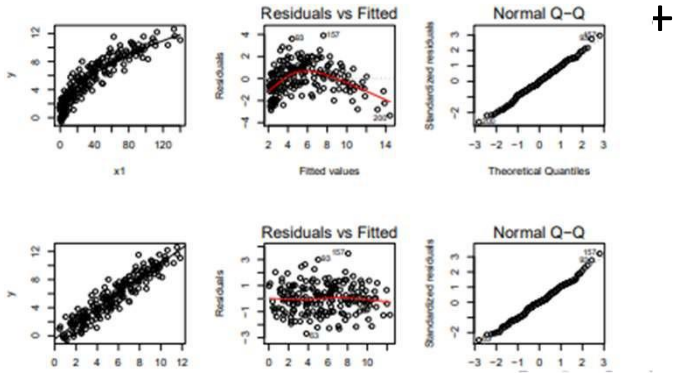
P-value	Decision	Conclusion
less than 5% ($p < 0.05$),	*Reject H_0*	*Significance*
equals to 5% ($p = 0.05$),	*Reject H_0*	*Significance*
more than 5% ($p > 0.05$),	*Fail to Reject H_0*	*Not Significance*

Transformations

- Transform response variable (y)
- Transform explanatory variable (x)(Square-root transformation of x, Power transformation of x, Inverse transformation of x).
- arithmetic transformation (e.g. log, inverse, squares etc)

Square-root transformation of x

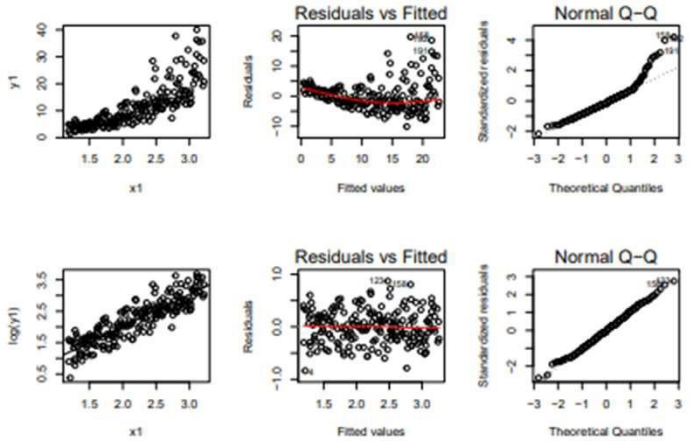
(e.g1) : Square-root transformation of x Original model $y = \beta 0 + \beta 1x1 + \varepsilon$.



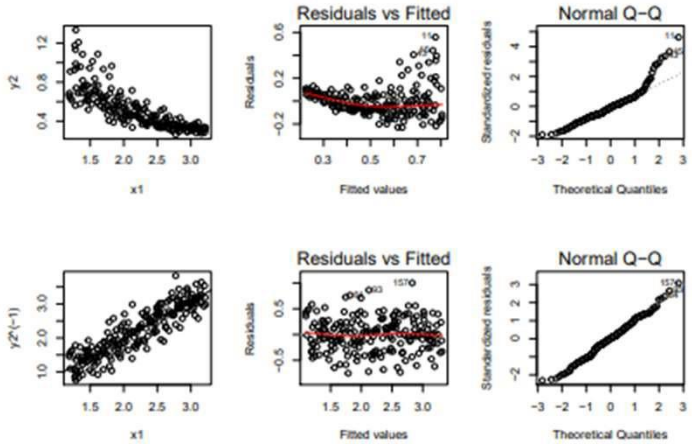
$+ \varepsilon$

Transformation of y

Log-transform of y



Reciprocal of y



Downside of transforming a variable is it complicates the interpretation of the model , sometimes it makes the data more skewed.

For example, in the absence of data transformation, y = β0 + β1x1 β1 is the increase in y when x1 increasing by one unit If we applied reciprocal transformation to y 1 y = β0 + β1x β1 becomes rate of decrease in y as x increases by one unit and It is important to back-transform the variable.

R-Markdown

R Markdown provides a notebook to:

- Save and execute code
  - Use an R Markdown file to load data, run analyses, connect to databases

Generate high quality reports to share with an audience

- Publish as a html, pdf, word file, slides, book, website etc...(Recommended is html and word as pdf has some issues).

Why use R Markdown

- Reproducible
- Readable (contains text + code)
- Share-able
- Easy to use with version control (*e.g.* git)
- ### References

R Markdown Reference-Additional details.
https://rmarkdown.rstudio.com/lesson-1.html)

Mandatory R -Supervised Learning techniques –Assignment Expectation

- Data Wrangling
- Exploratory Data Analysis
- Split the data-Train and Test (can be 70-30, or 80-20).
- Cross validation techniques- Most popular one K-fold- Determination of K value is key.
- Validation testing
- Hyper-Tuning
- Confusion Matrix
- Result Interpretation
- Results Discussions

References

https://www.stat.cmu.edu
https://datatab.net
https://www.geeksforgeeks.org/
https://github.com/MarthaCooper/
https://medium.com/
https://towardsdatascience.com/