LINEAR REGRESSION (QUESTION AND ANSWER)

- **What are the different types of Machine Learning?**

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	The machine learns by using labelled data	The machine is trained on unlabelled data without any guidance	An agent interacts with its environment by producing actions & discovers errors or rewards
Type of problems	Regression & Classification	Association & Clustering	Reward based
Type of data	Labelled data	Unlabelled data	No pre-defined data
Training	External supervision	No supervision	No supervision
Approach	Map labelled input to known output	Understand patterns and discover output	Follow trail and error method
Popular algorithms	Linear regression, Logistic regression, Support Vector Machine, KNN, etc	K-means, C-means, etc	Q-Learning, SARSA, etc

- **Explain Classification and Regression?**

Classification	Regression
• Classification is the task of predicting a discrete class label • In a classification problem data is labelled into one of two or more classes • A classification problem with two classes is called binary, more than two classes is called a multi-class classification • Classifying an email as spam or non-spam is an example of a classification problem	• Regression is the task of predicting a continuous quantity • A regression problem requires the prediction of a quantity • A regression problem with multiple input variables is called a multivariate regression problem • Predicting the price of a stock over a period of time is a regression problem

- **What is linear regression?**
- A linear regression is a linear approximation of a causal relationship between two or more variables.
- It falls under the supervised machine learning algorithms.

- **What is process of carrying out a linear regression?**
- Get sample data
- Design a model that works on that sample
- Make predictions for the whole population

- **How do you represent a simple linear regression?**
- Y = b0 +b1 x1 + e
- Y – dependent variable
- X1 – independent variable
- e – Error term = Y – Y(hat)

- **What is the difference between correlation and regression?**
- Correlation does not apply causation (the relationship between cause and effect). Regression is done to understand the impact of independent variable on the dependent variable.
- Correlation is symmetric regrading both the variables p(x,y) = p(y,x). Regression is one way.
- Correlation does not capture the direction of causal relationship. Regression captures the cause and effect.

LINEAR REGRESSION (QUESTION AND ANSWER)

- **What are the columns in the coefficient table?**
- The coefficient table contains the variable name, coefficient, standard error and p-value.

- **What is standard error?**
- Standard error shows the accuracy for each variable

- **What is p-value?**
- The p-value shows the significance of the variable. It tells us if the variable is useful or not.
- The H0 is coefficient = 0 and the H1 is coefficient ≠ 0
- If p-value < 0.05 (in most of the cases) we reject H0

- **What is OLS?**
- OLS stands for ordinary least square
- It measures the error between the actual Y and predicted Y
- Lower the error, better is the model

- **What are the other regression methods?**
- Generalized least squares
- Maximum likelihood estimates
- Bayesian regression
- Kernel regression
- Gaussian regression

- **What is TSS, ESS and RSS?**
- TSS stands for Total Sum of Squares. It measures the total variability (lack of consistency or fixed pattern).
- TSS = $\sum(y - y(mean))^2$
- ESS stands for Explained Sum of Squares. It measures the variability that is explained.
- ESS = $\sum(y(pred) - y(mean))^2$
- RSS stands for Residual Sum of Squares. It measures the difference between the observed Y and predicted Y.
- RSS = $\sum(y - y(pred))^2$

- **What is the relationship between TSS, ESS and RSS?**
- TSS = ESS + RSS
- Total variability = Explained variability + Unexplained variability

- **What is R-Squared?**
- R-Squared is also known as goodness of fit
- Smaller the RSS, better is the model
- R-Sq = ESS / TSS = 1 – (RSS / TSS)
- R-Squared takes a value between 0 and 1.

LINEAR REGRESSION (QUESTION AND ANSWER)

- If R-Sq = 0 then the model does not explain any variability
- If R-Sq = 1 then the model explains entire variability

- **What is adjusted R-Squared?**
- Adjusted R-Squared is a step on R-Squared and adjusts for the number of variables included in the model
- As we add more variables the explanatory power of the model may increase.
- Adjusted R-Squared penalizes the model for the number of variables that are used in the model.

- **What is the relationship between R-Squared and Adjusted R-Squared?**
- Adj R-Sq is always lower than the R-Sq
- Adj R-Sq = $1 - ((1-RSq) * (n-1) / (n-p-1))$
- Where n is the number of observations and p is the number of variables

- **What happens when we add a variable and it increases the R-Sq but decreases the Adj R-Sq?**
- The variable can be omitted since it holds no predictive power
- We should also look at the p-value of the added variable and confirm our decision

- **What is feature selection?**
- It is a method to simplify the model and improves the speed
- It is done to avoid too many features
- p-value in regression coefficient table can be used to drop insignificant variables

- **What is feature scaling?**
- Different variables have different magnitude
- Feature scaling is done to bring the variables to the same magnitude
- Standardization is one of the methods used for feature scaling

- **What is standardization?**
- It is also called normalization
- $X (std) = (x - \mu) / \sigma$
- Regardless of the data we will get data with mean 0 and standard deviation of 1

- **What is the interpretation of the weights?**
- In ML coefficients are called weights.
- A positive weight shows that as feature increases in value, so does Y
- A negative weight shows that as feature decreases in value, so does Y

- **What is the difference between overfitting and underfitting?**
- Underfitting happens when the model has not captured the underlying logic of the data.
- Overfitting happens when the model has focused too much on the training dataset that it cannot understand test dataset

- **How to identify if the model is overfitting or underfitting?**

LINEAR REGRESSION (QUESTION AND ANSWER)

- Underfit model performs bad (low accuracy) on training and bad (low accuracy) on test.
- Overfit model performs good (high accuracy) on training and bad (low accuracy) on test.
- A good model performs good (high accuracy) on training and good (high accuracy) on test.

- **What is multiple linear regression?**
- In multiple linear regression that are more than one predictor.
- Good models require multiple independent variables in order to address the higher complexity of the problem.
- $Y = b0 + b1\ x1 + b2\ x2 + \ldots + bk\ xk + e$

- **What are the assumptions of linear regression?**
- Linearity
- No endogeneity
- Normality and homoscedasticity
- No autocorrelation
- No multi-collinearity

- **What happens if the linear regression violates any of its assumptions?**
- The biggest mistake you can make is to perform a regression that violates one of its assumptions.
- If the regression assumptions are violated, then performing regression analysis will yield incorrect results.

- **What does linearity mean?**
- It means a linear relationship
- To check if there is linear relationship between x and y the simplest thing to do is plot a scatter plot between x and y

- **What are the fixes of linearity?**
- If linearity assumption is violated, then we can use non-linear regression
- We can also transform the x (exponential transformation / log transformation)

- **What does no endogeneity mean?**
- No endogeneity means no relationship between x and ε
- It may be because we have omitted an important predictor from the model

- **What is omitted variable bias?**
- If the modeler forgets to include an important predictor in the model
- It may lead to counter-intuitive coefficient signs
- Once the important variable is included rest of the coefficients fall into place

- **What is the assumption of normality?**
- It means the normal distribution of the error term
- The mean of the residuals should be zero
- The standard deviation of the residuals should be constant

LINEAR REGRESSION (QUESTION AND ANSWER)

- **What is the assumption of homoscedasticity?**
- In simple terms it means the equal variance
- There is no relationship between the error term and the predicted Y

- **How to prevent heteroscedasticity?**
- It may be due to outliers
- It may be due to omitted variable bias
- Log transformation

- **What does autocorrelation mean?**
- It is common in time series modeling
- It means that Y(t) is dependent on historical values, Y(t-1) or Y(t-2) or … Y(t-k)

- **How to detect autocorrelation?**
- DW (**Durbin Watson**) test is used to detect autocorrelation
- If DW test statistics is less than 1 then there is strong autocorrelation
- If DW test statistics is close to 2 then there is no autocorrelation
- If DW test statistics is more then 3 then there is strong autocorrelation

- **What are the remedies to remove autocorrelation?**
- There is no remedy in linear regression
- The modelers can try different models like AR, MA, ARMA or ARIMA

- **What does multicollinearity mean?**
- When two or more variables have high correlation
- If there is perfect multicollinearity then standard error will be infinite
- Imperfect multicollinearity means that the correlation is slightly less than 1 or slightly more than -1. However imperfect multicollinearity also causes serious issues in the model

- **What are the fixes of multicollinearity?**
- Find the correlation between each pair of independent variables
- If two variables are highly correlated, then either drop one of them or transform them into a single variable

- **What is VIF? How do you calculate it?**
- Variance Inflation Factor (VIF) is used to check the presence of multicollinearity in a dataset.

- **What are the disadvantages of the linear model?**
- Linear regression is sensitive to outliers which may affect the result.
- Over-fitting
- Under-fitting

- **How to find RMSE and MSE?**

LINEAR REGRESSION (QUESTION AND ANSWER)

- Answer???

- **What is the importance of the F-test in a linear model?**
- The F-test is a crucial one in the sense that it tests the goodness of the model. When you reiterate the model to improve the accuracy with the changes, the F-test proves its utility in understanding the effect of the overall regression.

- **How do you interpret a Q-Q plot in a linear regression model?**
- As the name suggests, the Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words, you plot quantiles against quantiles.
- Whenever you interpret a Q-Q plot, you should concentrate on the 'y = x' line. You also call it the 45-degree line in statistics. It entails that each of your distributions has the same quantiles. In case you witness a deviation from this line, one of the distributions could be skewed when compared to the other.

- **Pearson Vs Spearman correlation**
- **Pearson:** - The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.
- **Spearman**: - The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

- **Graph**
- **Bar graphs** to show numbers that are independent of each other.
- **Pie charts** to show you how a whole is divided into different parts.
- **Line graphs** show you how numbers have changed over time.
- **Cartesian graphs** have numbers on both axes, which therefore allow you to show how changes in one thing affect another. These are widely used in mathematics, and particularly in **algebra**.