Q1. Explain the linear regression algorithm in detail.

Answer: - Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price). Why Regression? Because the output variable to be predicted is a continuous variable, e.g. scores of a student.
Regression is the most commonly used predictive analysis model. It has its applications across variety of industries ranging from stock markets, finance, business to exit polls. As per CRISP DM framework before preparing predictive models, we need to first find out the business objective and do some data preprocessing. Linear Regression is a form of predictive modelling technique which tells us the relationship between dependent and independent variables.

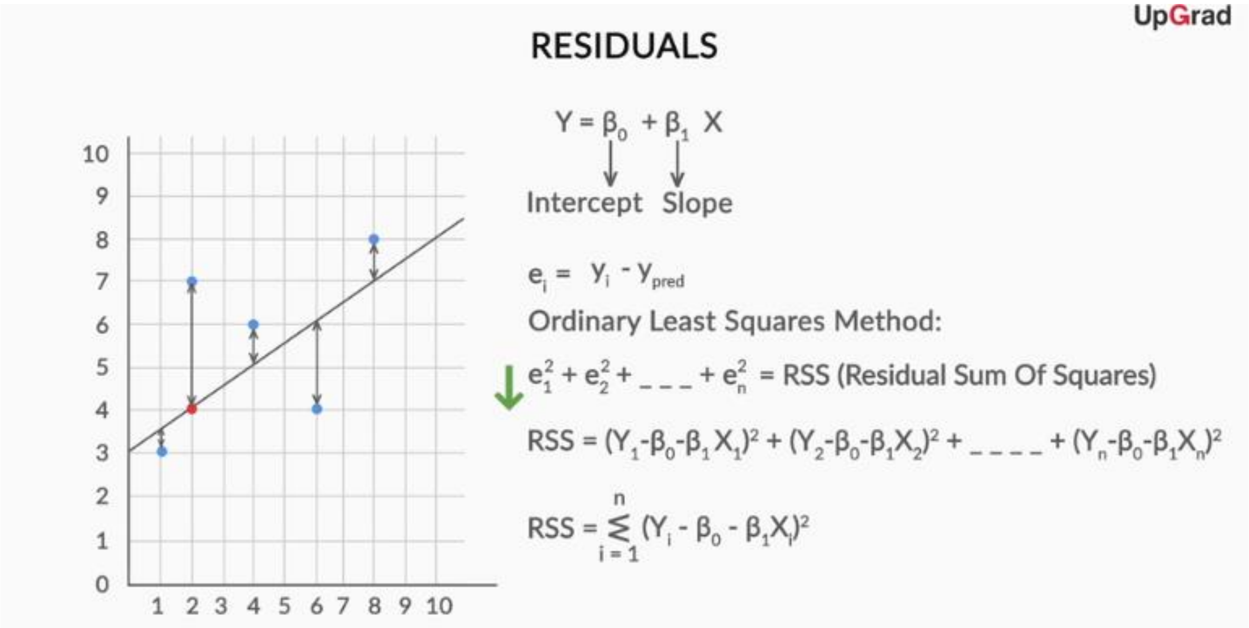
There are two main types of Linear Regression: -

1. **Simple Regression: -** The most elementary type of Linear Regression is Simple Linear Regression which explains the relationship between a target variable and a predictor variable using a straight line.

$$y = \beta_0+\beta_1X$$

y = Target Variable
X = Predictor Variable

- **Best Fit Line: -** The best fit line is obtained by minimizing the Residual sum of Squares which is equal to the sum of squares of residuals at each data point as shown below in diagram: -


UpGrad
RESIDUALS
$Y = \beta_0 + \beta_1 X$
Intercept
Slope
$e_i = y_i - y_{pred}$
Ordinary Least Squares Method:
$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)
$RSS = (Y_1-\beta_0-\beta_1X_1)^2 + (Y_2-\beta_0-\beta_1X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1X_n)^2$
$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1X_i)^2$
10
9
8
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10

- **Strength of Model assessed by two metrics: -**
R^2 – It is a number which explains what portion of the data variation is explained by the model. Mathematically,

$$R^2 = 1 –(RSS/TSS)$$

Here, RSS is Residual Sum of Squares and TSS is Sum of errors of data from mean

Overall, higher the R^2, the better the model

RSE(Residual Standard Error) – It is an estimate of the standard deviation of ε which is the error term or we can say that it explains on average, how is the prediction deviating from its original line. Mathematically,

$$\mathbf{RSE = (RSS/n\text{-}2)^{1/2}}$$

Here, n is sample size

- **Steps for creating a model: -**
  ***** Import the dataset and required libraries
  * Understand and prepare the data
  * Prepare predictor X and target y
  * Split the dataset into train and test
  * Performing Linear Regression either by using sklearn or statsmodel
  * Make prediction on test dataset using the model fitted on training dataset
  * Model Evaluation(Plot actual vs predicted and Error term)
  * Check mean squared error and R^2

2. **Multiple Linear Regression: -** Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables). The coefficient of the linear equation now measures the change in the target y, per unit increase in the variable when other predictors are held constant. Model now fits on hyperplane than on a line. Mathematically,

$$\mathbf{Y = \beta_0+\beta_1X_1+\ldots\ldots\ \beta_pX_p+\varepsilon}$$

The new aspects to consider from moving to linear regression to Multiple Linear Regression are: -

- **Dummy Variables**: - We need to convert categorical values to numeric form before regression modelling using pd.get_dummies()

- **Here Adjusted R^2** is the metric to be used instead of R^2 to assess how good fit is the model to the data because Adjusted R^2 penalizes R^2 for unnecessary addition of variables.

- **Multicollinearity**: - It may be possible that variables may be highly collinear with each other, this is called Multicollinearity. We can compute multicollinearity by Variance Inflation Factor. A variable greater than 5 needs to be removed as it show high collinearity with other variables.

- **Feature Scaling**: - The variables scale may vary from values ex -1,2 to 40k etc but we should scale them on a single scale. MinMaxScaler and StandardScaler can be used to scale the variables.

- **Variable Selection**: - Recursive Feature Elimination is based on the idea of repeatedly constructing model and then choosing either the best or the worst performing feature. The variables are ranked, we can then drop irrelevant variables based on ranking.

- **Model Selection**: -Build a model using either all variables or by choosing one by one. Check VIF and summary. Remove variables one by one by checking the significance (p>0.05) and VIF(>5). Repeat till we have all significant variables.

- **Model Evaluation**: - You can check model validation by checking the R^2 and RMSE of test data.

Q2. What are the assumptions of linear regression regarding residuals?

Answer: - There are five basic assumptions of Linear Regression Algorithm: -

- **Linear Relationship between the features and the target** – According to this assumption there is a linear relationship between the features and the target. Linear Regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target variable.

- **Little or No Multicollinearity –** It is assumed that there should be no or very little multicollinearity in data. Multicollinearity can be checked by below two criteria
  - **Correlation Matrix:** -We can check the correlation among independent variables by creating a correlation matrix, in python corr() gives back a correlation matrix.
  - **Variance Inflation Factor: -** VIF calculates how well one independent variable is explained by all the other independent variables combined. Mathematically,

    $$\mathbf{VIF = 1/(1-R_i^2)}$$

    With VIF > 5 there is an indication that multicollinearity may be present; with VIF > 10 there is certainly multicollinearity among the variables. However, it is assumed that VIF < 5 for linear regression.

- **Normal Distribution of error terms: -** The linear regression analysis requires all variables to be multivariate normal i.e. The error(residuals) follow a normal distribution. This assumption can best be checked with a histogram or a Q-Q-Plot.

- **Little or No Auto Correlation in Residuals: -** Linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other.

- **Homoscedasticity:** - Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a good way to check for homoscedasticity.

Q3. What is the coefficient of correlation and the coefficient of determination?

Answer: - Coefficient of Correlation - The quantity R, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson. The value of R is such that -1 < R < +1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively. The positive correlation between variables mean if one variable increases other increases as well while in negative correlation, when one increases the other decreases.

Coefficient of Determination: - The coefficient of determination, R^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph. The coefficient of determination is such that $0 < R^2 < 1$, and denotes the strength of the linear association between x and y. For example, if r = 0.922, then r 2 = 0.850, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

Correlation can be rightfully explained for simple linear regression – because you only have one x and one y variable. For multiple linear regression R is computed, but then it is difficult to explain because we have multiple variables involved here. That's why R^2 is a better term. You can explain R^2 for both simple linear regressions and also for multiple linear regressions.

Q4. Explain the Anscombe's quartet in detail.

Answer: - *Anscombe's Quartet* is the most elegant demonstration of the dangers of summary statistics. It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven *(x,y)* pairs as follows:

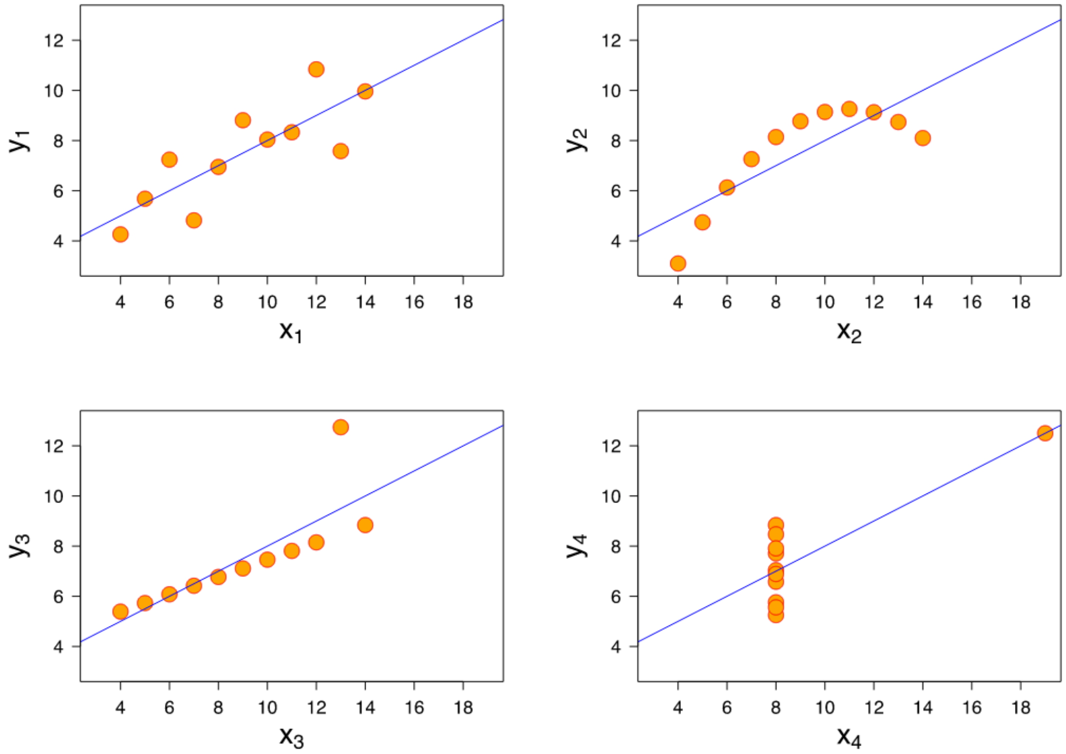
I		**II**		**III**		**IV**	
x	**y**	**x**	**y**	**x**	**y**	**x**	**y**
10.0	**8.04**	**10.0**	**9.14**	**10.0**	**7.46**	**8.0**	**6.58**
8.0	**6.95**	**8.0**	**8.14**	**8.0**	**6.77**	**8.0**	**5.76**
13.0	**7.58**	**13.0**	**8.74**	**13.0**	**12.74**	**8.0**	**7.71**
9.0	**8.81**	**9.0**	**8.77**	**9.0**	**7.11**	**8.0**	**8.84**
11.0	**8.33**	**11.0**	**9.26**	**11.0**	**7.81**	**8.0**	**8.47**

14.0	**9.96**	**14.0**	**8.10**	**14.0**	**8.84**	**8.0**	**7.04**
6.0	**7.24**	**6.0**	**6.13**	**6.0**	**6.08**	**8.0**	**5.25**
4.0	**4.26**	**4.0**	**3.10**	**4.0**	**5.39**	**19.0**	**12.50**
12.0	**10.84**	**12.0**	**9.13**	**12.0**	**8.15**	**8.0**	**5.56**
7.0	**4.82**	**7.0**	**7.26**	**7.0**	**6.42**	**8.0**	**7.91**
5.0	**5.68**	**5.0**	**4.74**	**5.0**	**5.73**	**8.0**	**6.89**

All the summary statistics you'd think to compute are close to identical:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

So far, these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:



Q5. What is Pearson's R?

Answer: -Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association

The formula for Pearson correlation coefficient r is given by:

$$\mathbf{R = n(\sum xy)-(\sum x)(\sum y)/ \surd\ [n\sum x2-(\sum x)2][n\sum y2-(\sum y)2]}$$

Where,
R = Pearson correlation coefficient
x = Values in the first set of data
y = Values in the second set of data
n = Total number of values.

Q6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: -We often handle datasets which had multiple features spanning varying degrees of magnitude, range, and units. For example, one feature is entirely in kilograms while the other is in grams, another one is liters, and so on. We can't use these features while building models when they vary so vastly in terms of what they're presenting. **The answer to Why we need to scale the variables in our dataset is** because some machine learning algorithms are sensitive to feature scaling while others are virtually invariant to it. Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled. Also, When you're working with a learning model, it is important to scale the features to a range which is centered around zero. This is done so that the variance of the features are in the same range. If a feature's variance is orders of magnitude more than the variance of other features, that particular feature might dominate other features in the dataset, which is not something we want happening in our model.

Standardized Scaling - Standardization (or Z-score normalization) is the process of rescaling the features so that they'll have the properties of a Gaussian distribution with $\mu=0$ and $\sigma=1$ where μ is the mean and σ is the standard deviation from the mean; standard scores (also called z scores) of the samples are calculated as follows:

$$\mathbf{Z = x- \mu/ \sigma}$$

Normalized Scaling - Normalization often also simply called **Min-Max scaling** basically shrinks the range of the data such that the range is fixed between 0 and 1 (or -1 to 1 if there are negative values). It works better for cases in which the standardization might not work so well. If the distribution is not Gaussian or the standard deviation is very small, the min-max scaler works better. Normalization is typically done via the following equation:

$$X_i = X_i - \min(X)/\max(X) - \min(X)$$

Q7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer - VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. If there is perfect correlation, then VIF is equal to infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation) **An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables** (which show an infinite VIF as well).

Q8. What is the Gauss-Markov theorem?

Answer - The **Gauss Markov theorem** tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible. There are five Gauss Markov assumptions (also called *conditions*):

- Linear Relationship between the features and the target
- Little or No Multicollinearity
- Normal Distribution of error terms
- Little or No Auto Correlation in Residuals
- Homoscedasticity

The **Gauss Markov assumptions** guarantee the validity of ordinary least squares for estimating regression coefficients.

Q9. Explain the gradient descent algorithm in detail.

Answer – Gradient Descent Algorithm is basically an optimization algorithm. **Optimization** refers to the task of minimizing/maximizing an objective function *f(x)* parameterized by *x*. In machine terminology, it's the task of minimizing the cost function. Its objective is to: -

- Find the global minimum of the objective function. This is feasible if the objective function is convex, i.e. any local minimum is a global minimum.
- Find the lowest possible value of the objective function within its neighborhood. That's usually the case if the objective function is not convex as the case in most deep learning problems.

Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost). Consider an example of a random position on the surface of the bowl is the cost of the current values of the coefficients (cost). The bottom of the bowl is the cost of the best set of coefficients, the minimum of the function. The goal is to continue to try different values for the coefficients, evaluate their cost and select new coefficients that have a slightly better (lower) cost. Repeating this process enough times will lead to the bottom of the bowl and you will know the values of the coefficients that result in the minimum cost.

While building linear regression model for the dataset we follow Gradient optimization algorithm where we continuously tend to approach to the strategies of choosing those features which tend to minimize the cost function. We repeat the process till we get the lowest possible cost function for a model.

For attaining a minimum cost function repeat the following steps until hit convergence:

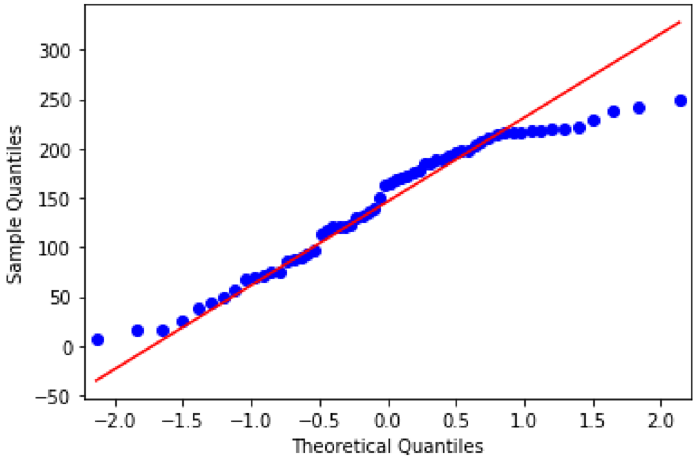
1. Given the gradient, calculate the change in the parameters with the learning rate.
2. Re-calculate the new gradient with the new value of the parameter.
3. Repeat step 1.

Q10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: -A quantile-quantile plot (also known as a QQ-plot) is another way you can determine whether a dataset matches a specified probability distribution. QQ-plots are often used to determine whether a dataset is normally distributed. Graphically, the QQ-plot is very different from a histogram. As the name suggests, the horizontal and vertical axes of a QQ-plot are used to show quantiles.

With a QQ-plot, the quantiles of the sample data are on the vertical axis, and the quantiles of a specified probability distribution are on the horizontal axis. The plot consists of a series of points that show the relationship between the actual data and the specified probability distribution. If the elements of a dataset perfectly match the specified probability distribution, the points on the graph will form a 45 degree line.

This plot shows if residuals are normally distributed. We can also check if residuals follow a straight line well or do they deviate severely. It's good if residuals are lined well on the straight dashed line. We can have a look at below graph that residuals follow a straight line for most of the path but deviates a little at the end points.


Sample Quantiles
Theoretical Quantiles