Linear Regression Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a linear model which aims to establish a linear relationship between the input variables (X) and a single output (y). When the input (X) is a single variable this model is called **simple Linear Regression** and when there are multiple input variables (X), it is called **multiple linear regression**.

We have two types of variables involved in linear regression:

- **Independent variables** (features): An independent variable is a variable that is manipulated to determine value of a dependent variable. These are features which we want to use to predict some value of (y). It is also called an explanatory variable.
- **Dependent variable** (target): The dependent variable depends on the values of the independent variable. It is the feature which we are trying to predict. Also called as outcome variable.

Linear regression is one of machine learning algorithms which can be used to predict some value (y) given a set of features (X).

If (y) is some outcome, and (X) some explanatory variable, then we can express the structural model using the equation

$$\mathbf{Y = \beta 0 + \beta 1 X}$$

where Y is “expected value of” indicating a population mean. It is read as indicating we are looking at the possible values of (y) when (X) is some single value.

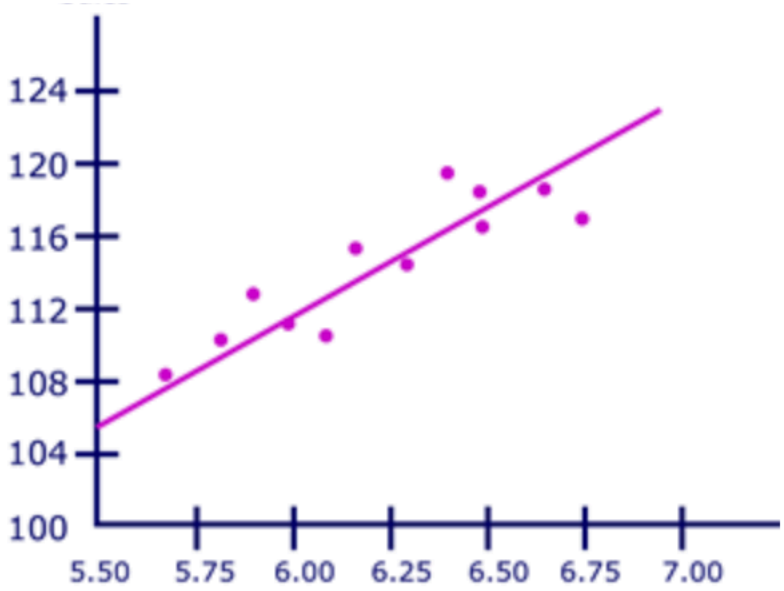
β0, “beta zero” is the **y-intercept** parameter; and β1, “beta one” is the **slope** parameter. These are **coefficients**.

For a multiple linear regression, the equation is like

$$\mathbf{E(y|X) = \beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots}$$

It explains the relationship between one continuous dependent variable (y) and two or more independent variables (X1, X2, etc.)

For simple linear regression,
the **null hypothesis** is H0: β1 = 0, and the **alternative hypothesis** is H1: β1 6= 0. If this null hypothesis is true, then we can say that the population mean of (y) is β0 for every (X) value, which tells us that (X) has no effect on (y). The alternative is that changes in (X) are associated with changes in (y) or changes in (X) causes changes in (y).



The **regression line** is the line with the smallest possible set of distances between itself and each data point. The distances of the data points from the regression line are called error terms. A regression line will always contain error terms because, in reality, independent variables are never perfect predictors of the dependent variables.

The typical procedure for finding the line of best fit is called the **least-squares method**. In this calculation, the best fit is found by taking the difference between each data point and the line, squaring each difference, and adding the values together. The least-squares method is based upon the principle that the sum of the squared errors should be made as small as possible, so the regression line has the least error.

Linear regression uses a **supervised "training and test"** set to predict a final result, predicated on their feature values. This is achieved by splitting data set into a training and test set. The purpose of the training set is to enable the machine to learn the relationship between independent variables and their respective target variable results. In doing so, we can then use the learned function to perform new predictions on un-labelled test set.

Algorithm steps followed in Linear Regression:

- Convert categorical variables to continuous ones using some **encoding** techniques like label encoding or derived variables.
- Create **dummy variables** or perform one hot encoding
- Perform **feature selection** to choose only relevant ones.
- Few techniques to **omit variables** are forward selection and backward elimination, or a mix of both.
- **Correlation** between multiple variables is also checked, because high multicollinearity affects the coefficient and accuracy and tends to increase the sum of squared errors.

Model evaluation interpretations are the following:

- If the **Prob (F-statistic)** is less than 0.05, we can conclude that the overall model fit is significant. If it is greater than 0.05, we might need to review the model as the fit might be by chance, i.e. the line may have just luckily fit the data.
- **R-squared** tells how much variance in the data has been explained by the model, and p-values of the coefficients tell us whether the coefficient is significant or not.
- **VIF** measure is also a good metric to detect multicollinearity. It's the ratio of variance of all coefficients and one variable's coefficient when it is the only variable in the model. VIF above 10 is considered to be misfit for the model.
- **Adjusted R squared** value measures how well the regression line fits with the actual results. Adjusted R squared penalizes for the increase in number of predictors.

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

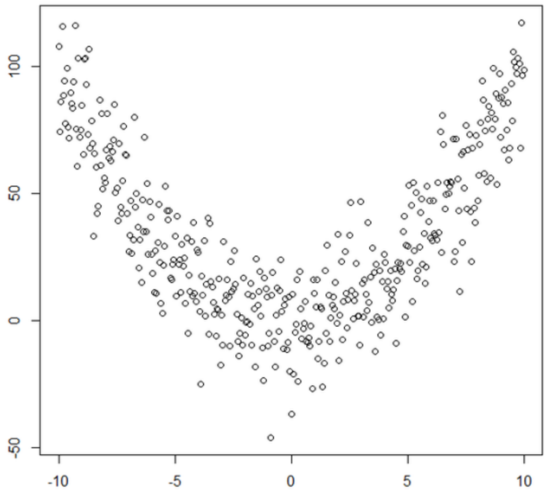
Answer:

In regression analysis, the difference between the observed value of the dependent variable (y) and the predicted value (ŷ) is called the residual (e). Each data point has one residual.

Residual = Observed value - Predicted value

An important way of checking whether a regression, simple or multiple, has achieved its goal to explain as much variation as possible in a dependent variable while respecting the underlying assumption is to check the residuals of a regression. Following are the assumptions of linear regression regarding residuals.

a) **Linearity**:
The regression model is having linear relationship between the features and target variable. If the scatter plot between two variables follows a linear pattern that shows linearity assumption is met. For e.g., below plot is non linear.



b) Mean of residuals (errors) is zero
We should check mean of all residual values and if it comes out to be zero, or close to zero, then this assumption holds true for the model. Also, residuals should not be skewed.

c) **Error term (residuals) should be normally distributed**

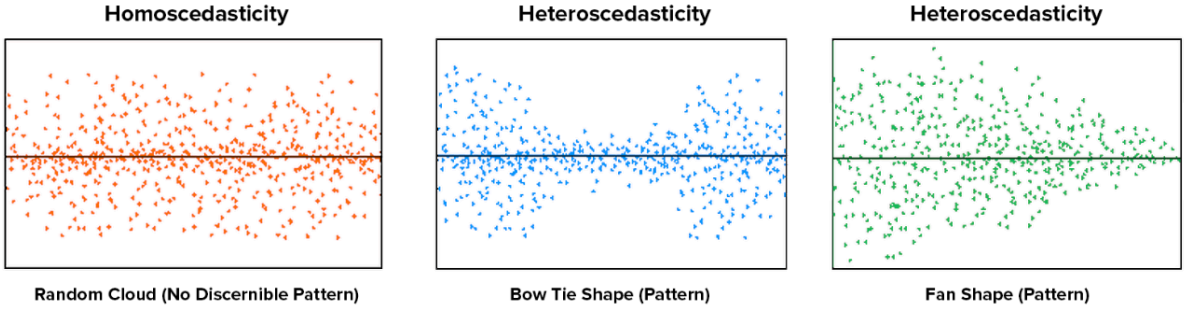
Null hypothesis is set to that residual is normally distributed. If the p-value is greater than 0.05, it means we cannot reject the null hypothesis that residual is normally distributed.

d) **No multicollinearity between features**

There must be no or very less correlation among independent variables. Presence of correlation in independent variables lead to Multicollinearity. If variables are correlated, it becomes extremely difficult for the model to determine the true effect of independent variables on dependent/target variables.

e) **Homoscedasticity of residuals or equal variance**

The variance of residuals should be approximately equal for all values of independent variables. If not, leads to heteroscedasticity where the estimators are no longer the Best Unbiased Estimators. The hypothesis tests (t-test and F-test) are also no longer valid.



f) **Error terms must be uncorrelated**

The error terms must be uncorrelated i.e. error at X value (t) must not indicate the at error at X value (t+1). Presence of correlation in error terms is known as Autocorrelation. It drastically affects the regression coefficients and standard error values since they are based on the assumption of uncorrelated error terms.

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

Answer:

The **correlation coefficient**, denoted by r, is a measure of the strength of the straight-line or linear relationship between two variables. The correlation coefficient takes on values ranging between +1 and -1. It is the degree to which two sets of numbers go together.

Col A	Col B	Col C	Col D
9	6	1	6
8	5	2	2
7	4	3	8
6	3	4	1
5	2	5	5
4	1	6	3
3	0	7	9

- Correlation coefficients can be highly positive as +1.00 if the relationship between the two sets of numbers is perfect and in the same direction (e.g. in table above for Col A and Col B).
- It can even be too high in a negative direction as -1.00 if the relationship is perfect and in opposite directions (e.g. in table above Column A and Column C, or Column B and Column C).
- It's zero if no relationship at all exists between two sets of numbers e.g. random numbers

The correlation coefficient requires that two variables have linear relationship. If nonlinear then the it's is not much useful. The correlation coefficient is defined as the mean product of the paired standardized scores (zXi, zYi) as expressed below:

zXi = [Xi - mean(X)]/std(X)
zYi = [Yi - mean(Y)]/std(Y)

r (X, Y) = sum of [zXi * zYi]/(n-1), where n is the sample size

The **coefficient of determination** is the measurement of how close the data are from the fitted regression line. It is simply the squared value of the correlation coefficient. The resulting coefficient of determination provides an estimate of the proportion of overlapping variance between two sets of numbers. It's also known as R squared.

R square is measured on a scale of both 0 to 100 and 0 to 1, with a measurement of 100 indicating that the dependent/target variable performance is entirely determined by an independent variable. A low R square indicates that there is no significant relationship between them.

For example, if we have two sets of scores on Tests X and Y, and they correlate at .90, we can square that value to get the coefficient of determination of .81 and interpret that result as meaning that 81% of the variance in Test X is shared with Test Y, or that 81% of the variance on Test Y is shared with Test X. So, the coefficient of determination is the proportion of the explained variation relative to the total variation.

If the points are close to a straight line, then the unexplained variation will be a small proportion of the total variation in the values of the response variable. This means that the closer the coefficient of determination is to 1 the stronger the linear relationship.

It is expressed as **R square = SSR / SST**,
where SSR is the sum of square of residuals, residual is the difference between the predicted value and the mean value.
SST is the total sum of squares. It is calculated by summing the squares of difference between the actual value and the mean value.

Here, SST = SSR + SSE,
where SSE is the sum of the square of Error
It is calculated by summing the square of the difference between the actual value and predicted value.

4. Explain the Anscombe's quartet in detail.

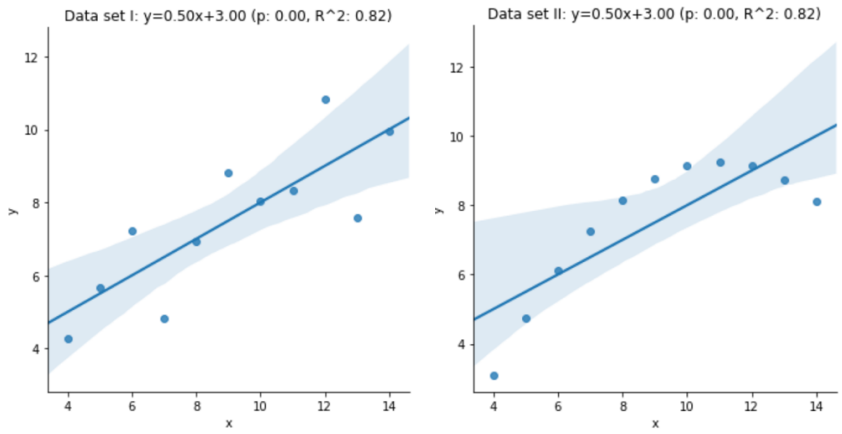
Answer:

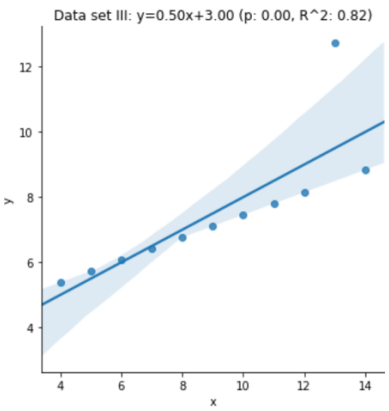
Anscombe's quartet is a **collection of four datasets** that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (X, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance and convenience of graphing data & data visual representation before analyzing it.

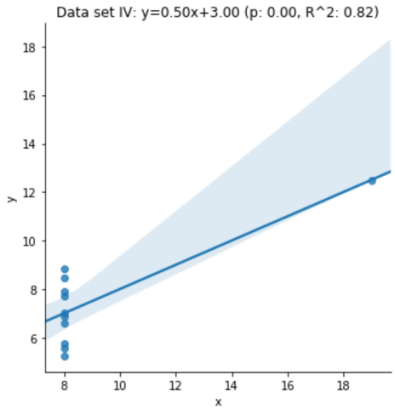
Each of these datasets seems to be quite different. But they are quite similar statistically speaking. Each of the four plots has the same mean and variance on both axes, as well as the same correlation coefficient, & same linear regression. Details as below:

	x								y							corr
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	
dataset																
I	11.0	9.0	3.316625	4.0	6.5	9.0	11.5	14.0	11.0	7.500909	2.031568	4.26	6.315	7.58	8.57	10.84	0.82
II	11.0	9.0	3.316625	4.0	6.5	9.0	11.5	14.0	11.0	7.500909	2.031657	3.10	6.695	8.14	8.95	9.26	0.82
III	11.0	9.0	3.316625	4.0	6.5	9.0	11.5	14.0	11.0	7.500000	2.030424	5.39	6.250	7.11	7.98	12.74	0.82
IV	11.0	9.0	3.316625	8.0	8.0	8.0	8.0	19.0	11.0	7.500909	2.030579	5.25	6.170	7.04	8.19	12.50	0.82

- The mean X value is 9 for each dataset
- The mean y value is 7.50 for each dataset
- The variance for X is 3.32 and the variance for y is 2.03
- The correlation between x and y is 0.82 for each dataset

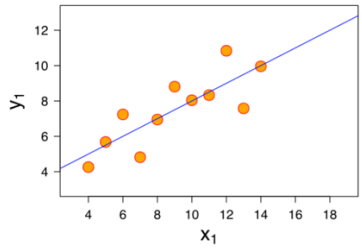


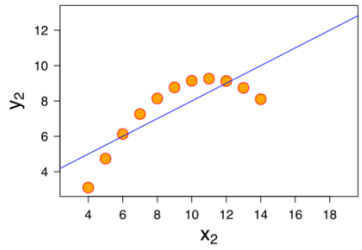


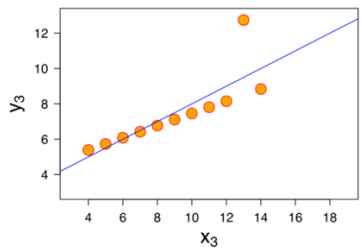


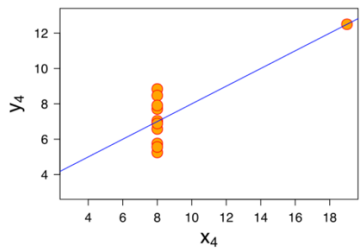
A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3. So far, these four datasets appear to be pretty similar statistically.

But, when we plot these four datasets on an X/y coordinate plane, can immediately spot the differences between them when we visualize them. We observe that they show the same regression lines as well, but **each dataset is telling a different story**.









- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

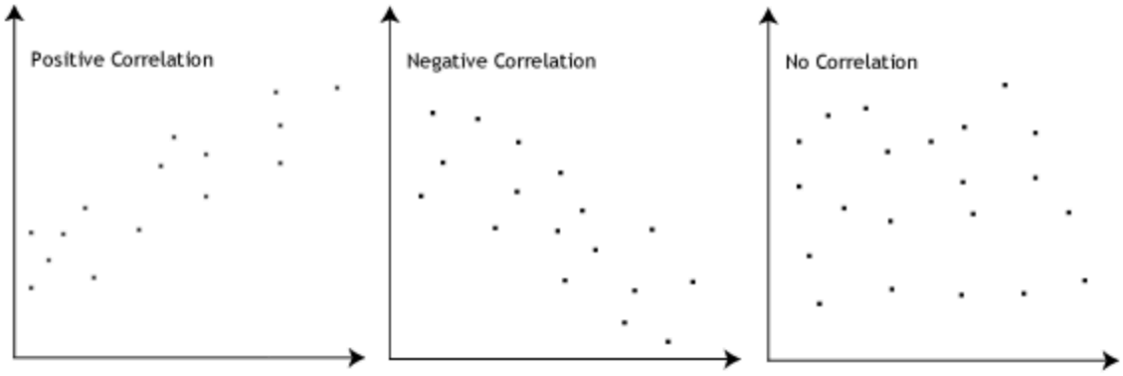
Thus, Anscombe's quartet is often used as an example justifying that a summary of a data set will inherently lose information and so should be accompanied by further visual study/understanding.

5. What is Pearson's R?

Answer:

Pearson's r is referred to as Pearson's correlation or simply as correlation coefficient. It denotes the **strength of the linear association between the variables**. If the variables tend to go up and down together, it will be positive. If the variables tend to go up and down in opposition, it will be negative.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Most widely used representation is r and it can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship, 0 means no linear relationship & r of 1 is a perfect positive linear relation.



The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

The Pearson's r correlation does not take into consideration whether a variable has been classified as a dependent or independent variable. It treats all variables equally.

Pearson correlation coefficient, r, does not represent the slope of the line of best fit. Therefore, if you get a Pearson correlation coefficient of +1 this does not mean that for every unit increase in one variable there is a unit increase in another. It simply means that there is no variation between the data points and the line of best fit.

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

Answer:
Scaling is used on datasets which contain features varying highly in magnitude, units or range. It is **used to unify feature ranges in data**. Since linear regression uses Euclidian distance between two data points, we should suppress this effect by bringing all features to the same level of magnitudes. This can be achieved by scaling.

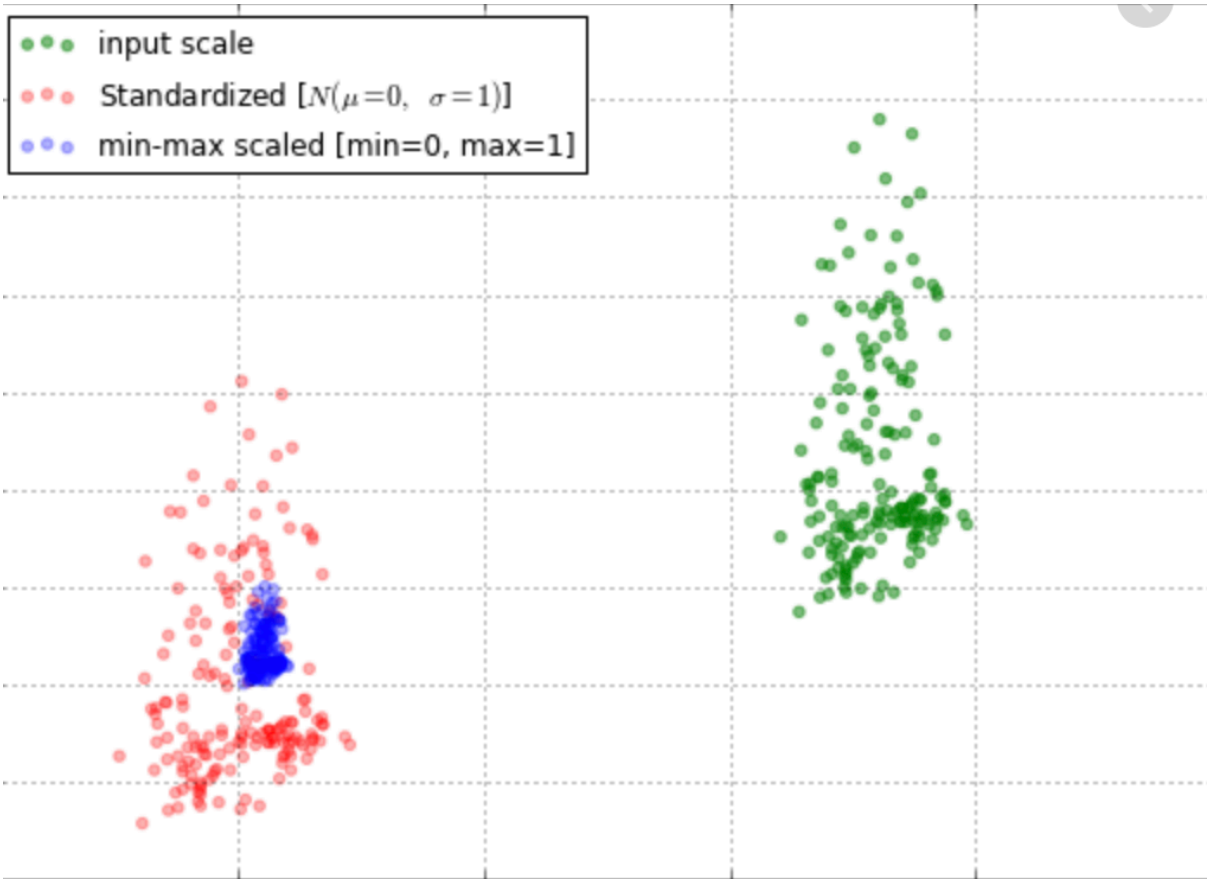
Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc. Among many ways, there are two most prominent methods for feature scaling: normalization method and standardized scaling. These are just different hyperparameters to give similar results.

Normalized scaling is achieved by subtracting the minimum value for datapoint in a feature followed by dividing by the adjusted maximum value for that feature values. This scaling happens in such a way that the smallest value for each feature becomes zero and the largest value becomes one, thus also known as 'Min. to Max. Scaling'. It scales all of the values for each feature so that they all fall in the range from 0 to 1. Equation is:

$$x = (x-\min(x)) / (\max(x) - \min(x))$$

Standardization scaling performs normalization by subtracting the mean and dividing the standard deviation for each feature. It basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. The mean and standard deviation for each feature is calculated, and each value for the variable (X) is standardized using (X -mean)/standard deviation. The formula used:

$$x = (x-\text{mean}(x)) / \text{standard deviation}(x)$$



We can see that our actual data (in green) is spread between higher units of the plot, **standardized data (in red)** is spread around the lower units of plot whereas **normalized data i.e. min. to max. scaling (in blue) i**s spread around 0 to 1 unit only.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers. For clustering, standardization may be quite crucial in order to compare similarities between features based on certain distance measures.

To use these, there is a convenient approach using the preprocessing module from one of Python's open-source machine learning library scikit-learn. To use min-max scaling technique, we can import MinMaxScaler from sklearn. preprocessing module or to use the standardization, we can import StandardScaler. Fit and transform methods are used to perform this action.

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

Answer:

Machine learning algorithms deal with huge dimensions of data and it's challenging to build simplest linear model possible. In this process of building the model there are possibilities of two or more predictor variables highly correlated to each other.

X1	X2	X3	y
1	2	2	5
2	3	3	8
3	4	4	11
4	5	5	14
5	6	6	17
6	7	7	20
7	8	8	23

If asked to write a linear equation for the above dataset with the basic math skills, we can come up with following linear combinations of variables.

S.no	Equation
1.	y = X1 + X2 + X3
2.	y = X1 + 2 X2
3.	y = 3 X1 + 2

Of these three possibilities, (3) is better to choose as we can compute the value of y with only one variable x. But it also means that X1, X2, and X3 are highly correlated variables. Correlation coefficient will clearly show that all these three columns are positively & highly correlated and hence it is viable to use only one variable instead of all three.

It's a challenge for basic machine learning models to identify automatically these highly correlated variables and pick only one of these.

To identify highly correlated variables. one of the most commonly used methods is Variation Inflation Factor (VIF). A VIF value of greater than 10 means that the variable is more than 90% correlation with other variables in the dataset and we can choose to remove these variables and re-run the model. To use VIF, we can import variance_inflation_factor from outliers_influence module featured in statsmodels. stats library.

VIF value of "inf" is infinity as the correlation is 1 exactly. If the relation between the independent variables grows really strong then the variance of parameter estimates tends to be infinity. So, we can remove such combination of variables from the data, pick just one among such highly correlated variables and rebuild the model. In a way, special care should be taken when the value of VIF is shown as inf for a variable.

Few advantages of performing such collinearity study and removing highly collinear variables are the following:

- By treating multicollinearity in the variables, machine learns to build possibly simple predictive model with minimum variables we pass.
- For larger data set removing highly collinear features will help to build models relatively faster.
- Model would become sensitive with highly collinear variables, which means that with small variations in the data, model would return large residual errors in the prediction.

8. What is the Gauss-Markov theorem?

Answer:

Gauss-Markov theorem applies to a Linear Regression model with random sampling. Following are some concepts leading us to the theorem, using example with just one explanatory variable:

$$y(i) = \beta 0 + \beta 1 x(i) + e, \text{ where } i = 1, \ldots, n$$

- In linear regression model, an estimator is linear if it is a linear function of y1, …, yn. Then OLS (Ordinary Least Square) estimators β0, β1 are also linear estimators.
- An unbiased estimator is more efficient than another unbiased estimator if it has a smaller variance.
- An estimator is the best linear unbiased estimator (BLUE) if it is linear and unbiased and more efficient than any other linear and unbiased estimator

Gauss-Markov states if we have two different estimators and both are best linear and unbiased estimators with same variance.

- In Linear Regression Model, where the errors are uncorrelated, have equal variances and expectation value of zero,
- **the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, if exists**. It should be the lowest variance of the estimate, as compared to other unbiased, linear estimators.
- The errors do not need to be normal, nor do they need to be independent and identically distributed. They should only be uncorrelated with mean zero and same finite variance.
- The requirement that the estimator be unbiased cannot be dropped, since biased estimators might exist with lower variance.

OLS estimator is a BLUE because the least-squares estimator is uncorrelated with every linear unbiased estimator of zero, i.e., with every linear combination whose coefficients do not depend upon the unobservable, but whose expected value is always zero.

9. Explain the gradient descent algorithm in detail.

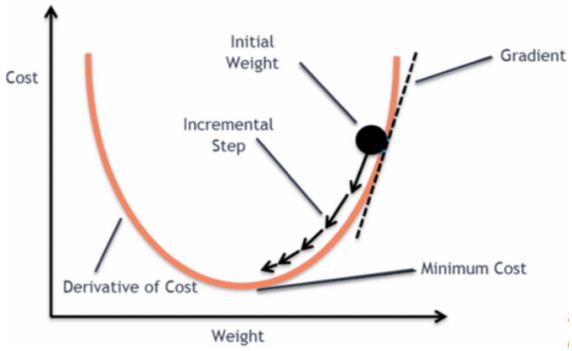
Answer:

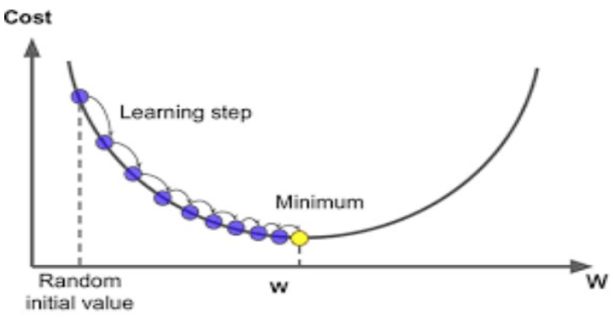
A **cost function** is a mechanism utilized in supervised machine learning to return the error between predicted outcomes compared with the actual outcomes. The aim of supervised machine learning is to minimize the overall cost, thus optimizing the correlation of the model to the system that it is attempting to represent.

The gradient descent algorithm aims to **minimize a cost function** of a model in order to produce a model that gives the most accurate predictions. In mathematics, the gradient of a function represents the steepness of said function. Gradient descent is the most commonly used algorithm in machine learning for this.

Working of gradient descent algorithm:

- Descends to the lowest point in an n-dimensional space.
- The lowest point represents the lowest value, and hence the minimum value of the cost function.
- Gradient descent uses the derivative of the function (the gradient of the curve) and changes the parameters of the algorithm in small steps (determined by a learning rate), each time moving towards another point that has a smaller cost.
- Eventually, gradient descent converges to a point where the gradient is close to 0, the minimum value for the cost function.
- The parameters that this algorithm yields will form the best model as it produced the minimum error in prediction.





Gradient Descent Implementation steps:

- Initialize parameters (weight and bias) with random value or simply zero. Also initialize Learning rate.
- Calculate Cost function by computing the loss error comparing the expected output and predicted value.
- Take Partial Derivatives of the cost function with respect to weights and biases.
- Update parameter values of weight and bias with learning rate.
- Repeat step2 to step4 till n no. of iterations. With each iteration the value of cost function will progressively decrease and eventually become flat value.

When a machine is trying to learn how to do a task, it tries to come up with an algorithm that makes as few mistakes as possible i.e. minimize its error function. The algorithm trains over all the training examples it knows. Then, it uses the gradient of the error function to adjust the configuration of the algorithm to reduce the total error over all training examples. It repeats this process over and over until it can no longer reduce the error.

There are different types of Gradient Descents defined on the basis of how we use the data to calculate the cost function derivative.

- **Batch Gradient Descent**: Whole training data set is checked upon for any errors and then it is updated back to compute the gradient of the cost function.

- **Stochastic Gradient Descent**: This approach performs multiple epochs to get the best results. One epoch is when all inputs from a training set are processed. It keeps modifying the weights after each processed input of the training set.

- **Mini-Batch Gradient Descent**: It's a mix of above approaches and uses mini batches of instances to update the model and to get the best results.

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

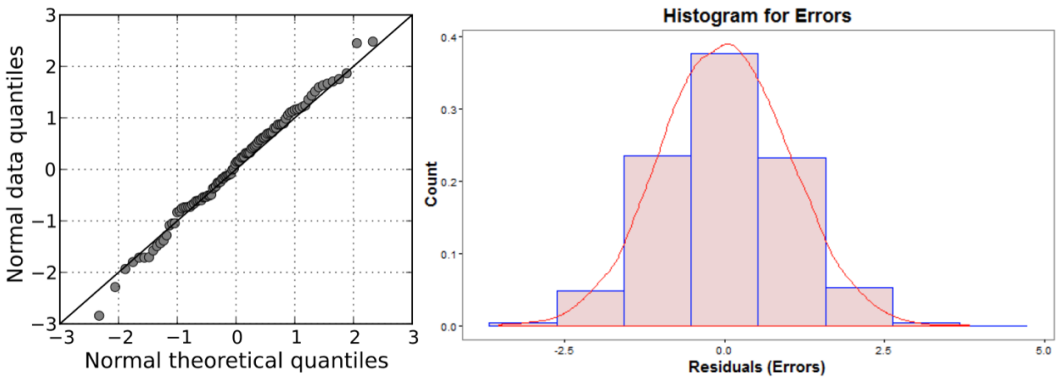
Answer:

A quartile-quartile plot is famously known as Q-Q plot. This plot charts the **quantiles of an observed sample against the quantiles from a theoretical normal distribution**.

The more linear the plot, the more closely the sample distribution matches the normal distribution. It helps to verifies the normality of errors of a regression model. It's also known as Quantile Comparison, Normal Probability, or **Normal Q-Q plots**.

To construct the normal Q-Q plot three steps are followed:

- Arrange residuals in ascending order.
- Find the Z-scores corresponding to N+1 quantiles of the normal distribution, where N is the number of residuals.
- Plot the sorted residuals on the vertical axis and the corresponding z-score on the horizontal axis.



If the observed distribution of the residuals matches the shape of the normal distribution, then the plotted points should follow a 1-1 relationship. If the points, follow the displayed straight line that suggests that the residuals have similar shape to normal distribution.

This plot also helps in analyzing the tail of distributions. Heavy-tailed residual distributions can be problematic for models as the variation is greater than what the normal distribution can account for and our methods might under-estimate the variability in the results.