1. Explain the linear regression algorithm in detail.

Linear regression algorithm is a type of supervised learning algorithm. Let us break down the term "Linear Regression" in order to better understand it –

- Linear models describe a straight-line relationship between variables.
- Regression is a type of supervised learning method where the output of a predictive analysis is a continuous variable. It is also the most commonly used predictive analysis model.

Therefore, the goal of Linear Regression algorithm is to establish a straight line (linear) relationship between a continuous output / target variable with a predictor or independent variable. There are multiple straight-line relationships that can be formed from a set of two variables and therefore, the goal, to be more specific, is not just to establish a linear relationship, but the "best-fit line" relationship between the output and the predictor.

Depending on the number of predictor variables linear regression algorithms can be divided into two:

1. Simple Linear Regression

   The simple Linear Regression explains the relationship between a dependent variable and one independent variable using a straight line. The standard equation of the regression line is given by the slope intercept equation in coordinate geometry (y = mx +c) as:

$$Y = \beta_0 + \beta_1 X;\ \ \beta_{\,0}\ = intercept\ and\ \beta_1 = slope$$

   The best fit line is obtained by minimizing the Residual Sum of Squares (RSS), which is the sum of squares of the residuals, where residuals are the difference between predicted value and the actual value of the dependent variable. This technique is known as Ordinary Least Squares Method (OLS).

$$e_i = y_i - y_{pred}$$

$$RSS = (Y_1 - \ \beta_0 - \ \beta_1 X_1)^2 + \ (Y_2 - \ \beta_0 - \ \beta_2 X_2)^2 + \cdots (Y_n - \ \beta_0 - \ \beta_n X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \ \beta_0 - \ \beta_i X_i)^2$$

2. Multiple Linear Regression

   Multiple Linear Regression is a statistical technique that explains the relationship between one dependent variable and multiple independent variables. The objective therefore is to find a linear equation that can best determine the value of dependent variable Y for different values of independent variable X_1-X_n. The linear model is very similar to that of simple regression model however with a few assumptions in place.

$$Y = \beta_0 + \beta_1 X_1 + \ \beta_1 X_2 + \cdots \ \beta_n X_n;\ \beta_{\,0}\ = intercept$$

$$\beta_i is\ the\ change\ in\ value\ of\ Y\ for\ 1\ unit\ of\ change\ in\ X_i\ when\ other\ X\ variables\ are\ kept\ constant$$

.

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

   The basic assumption of Linear Regression model is that the relationship established by dependent and independent variables is linear or approximately linear. Other assumptions are about error terms (residuals)

   a. The error terms are normally distributed with a mean of 0.

      If only a best fitted line is required, and no further interpretations are being made on the predictors or model it would be acceptable to have error terms that are not normally distributed. A repercussion of error terms not being normally distributed is that the p values obtained during hypothesis testing to determine the significance of coefficients become unreliable. This is common when the error distribution is skewed by outliers. Since parameter estimation is based on the minimization of mean squared error, a few extreme observations can exert a disproportionate influence on parameter estimates.

   b. The error terms are independent of each other.

      For instance, dependence of error term on previous value such as time dependence between consecutive errors exhibited by time series data. The violation of this assumption indicates that there can be improvements made on the model. Another example would be if error tends to always have the same sign, indicating that the model is systematically underpredicting or overpredicting.

   c. Error terms have constant variance (homoscedasticity)

      The variance should neither increase nor decrease as error value changes. The variance should not follow any specific pattern with error values either. If this assumption is violated it will result in confidence intervals that are too wide or too narrow. If the variance of the errors increases over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. Heteroscedasticity may also have the effect of assigning too much weight to a small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.

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

   Coefficient of Correlation: Is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A correlation of -1.0 shows a perfect negative correlation – which implies that increase in one variable will cause an equivalent decrease in the second variable. While a correlation of 1.0 shows a perfect positive correlation, which implies that an increase in one variable will cause an equivalent increase in the other as well. A correlation of 0.0 shows that there is no discernible relationship between the movement of the two variables. There are several types of coefficients but the commonly used is Pearson correlation (R), which measures strength and direction of the linear relationship between two variables. It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

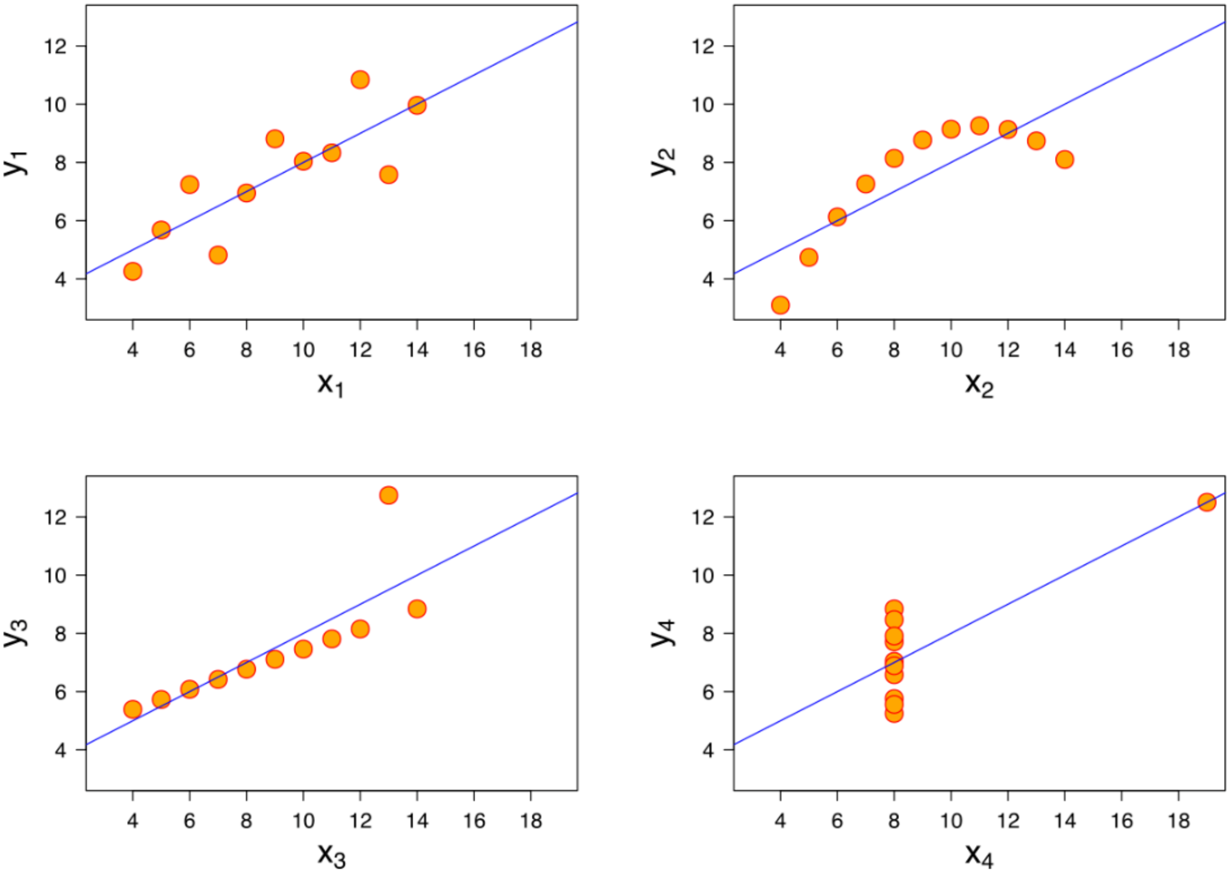
   Coefficient of Determination: Is used to assess the strength of linear regression model. It is denoted by R^2 and represents the proportion of variance in the output variable that is predicted from the independent variable. The value varies between 0 (indicating that the variance in output is not explained by the independent variable at all) and 1 (indicating the variance in output is entirely explained by the independent variable). Simply stated, it provides a measure of how well the actual outcomes are replicated by the model. Higher the R^2, better the model fit of data.

$$R^2 = 1 - \frac{RSS}{TSS};$$
$$RSS = Residual\ sum\ of\ squares$$
$$TSS = Residual\ sum\ of\ errors\ of\ data\ from\ mean$$

The worst possible linear regression model would be a line passing through mean of the target variables. TSS therefore gives the deviation of all points from the mean line.

In the specific case of linear regression, the coefficient of determination is equal to the square of the correlation between x and y scores. That is, Pearson's correlation R is the coefficient of correlation and R^2 (the coefficient of determination) is the square of Pearson's correlation R.

4. Explain the Anscombe's quartet in detail.



The above 4 sets are identical when examined using modelling and simple summary statistics but vary considerably when graphed. They were constructed to demonstrate the importance of visualizing data before analyzing it and the effect of outliers and other influential observations on statistical properties. It was intended to counter the impression that numerical calculations are exact, but graphs are rough. The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

For all 4 of the above graphs the mean of x, mean of y, variance of x, variance of y, correlation between x and y, linear regression line and coefficient of determination of linear regression are equal or approximately equal.

- The first (top left) scatter plot represents a simple linear relationship between 2 variables with assumption of normality.
- The second graph (top right) is not distributed normally and though the variables are related, the relationship is non-linear and therefore the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from a perfect fit of 1 to a 0.8 R^2
- The fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

5. What is Pearson’s R?

Pearson's correlation coefficient (r) is a measure of the strength of the association between two quantitative, continuous variables, for example, age and blood pressure. The correlation coefficient is not relevant for non-linear relationships. When a scatter plot is plotted between the independent variable and the target variable, nearer the plots are to a straight line, higher the strength of association (R) of the variables. The measurement unit of the variables themselves do not affect the Pearson’s correlation coefficient.
Pearson's correlation coefficient (R) for continuous (interval level) data, Cauchy–Schwarz inequality, ranges from -1 to +1. A value of 0 indicates no linear correlation. Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, the other decreases, and vice versa.

$$\rho\, x, y\ = \frac{cov\ (x, y)}{\sigma x\ \sigma y}$$

Pearson’s correlation coefficient is the covariance of the two variables divided by the product of their standard deviation.

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

Scaling in linear regression is the process of making the coefficients of predictors (parameters) comparable. This is importance especially in the case of the predictors having disparate scales, say one variable is a binary value whereas the other is in millions. The reason for performing scaling is because the coefficients of a scaled model where the values of all the columns are comparable, would be solely based on the significance of the predictor itself rather than on the scale of the predictor.

The null hypothesis for linear regression is that the value of $\beta_i = 0$. If there are 2 variables X1 with binary values and X2 with value in millions and the target variable too has value in millions, chances are that the coefficient of X1 would be so skewed due to different scales of the variables that it may become 0. This results in a Type 2 error where we fail to reject the null hypothesis even though it is false.

Scaling only affects the coefficients and does not affect other parameters like t-statistics, F-statistics, p-value, R^2.

The two common ways of rescaling are:

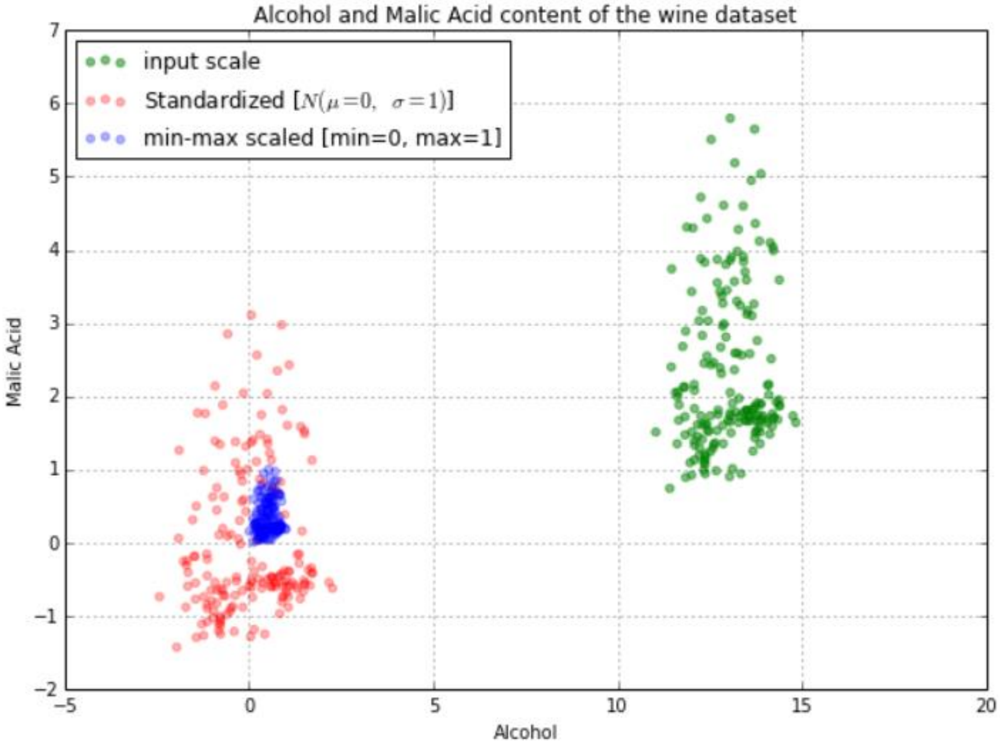
1. Min Max Scaling: This type of scaling rescales the entire set of values such that the minimum value for a column is always 1 and the maximum value is always 0. Min Max scaling brings all data in the range of 0 and 1. Min Max scaling compresses the entire dataset range of a column within the small range of 0-1.

$$x = \frac{x - \min(x)}{max(x) - \min(x)}$$

2. Standardization: This type of scaling converts the data to a standard normal distribution with mean 0 (centered at 0) and standard deviation of 1. The advantage of standardization over min max scaling is that data is not compressed. The entire dataset is simply shifted.

$$x = \frac{x - \text{mean}(x)}{sd(x)}$$

A pictorial comparison between the change in values caused by the min max and standardization scaling is given below.



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

Variance Inflation Factor is a statistic measure used to assess multicollinearity between predictor variables while performing multiple linear regression. Higher the VIF, higher the multicollinearity, implying the variable can be largely explained by other independent variables.
The VIF is given by:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where 'i' refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.
VIF can become infinite if R2 = 1, meaning if the variables are perfectly correlated. If an independent variable can be perfectly explained by a linear combination of other independent variables, the VIF for that independent variable will be infinite. There is no upper bound to VIF values.

8. What is the Gauss-Markov theorem?

Gauss Markov theorem applies for a linear regression model in which

- errors are correlated
- variances are equal
- expectation value is 0

The Best Linear Unbiased Estimator (BLUE) of the coefficients of the above model is given by the Ordinary Least Squares (OLS) estimator, if it exists. The best model is determined as one that gives the lowest variance of the estimate in comparison to other biased linear estimators. The errors need to be neither normal nor independent and identically distributed.

There are five Gauss Markov assumptions (also called conditions):

- Linearity: the parameters we are estimating using the OLS method must be themselves linear.
- Random: our data must have been randomly sampled from the population.
- Non-Collinearity: the regressors being calculated aren’t perfectly correlated with each other.
- Exogeneity: the regressors aren’t correlated with the error term.
- Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients. Verifying how well data matches these assumptions is an important part of estimating regression coefficients. In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they show us what ‘ideal’ conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by
Algebraically

$$y_i = x_i{}' \beta + \varepsilon_i$$

and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

$E\{\varepsilon i\} = 0, i = 1, \ldots, N$
$\{\varepsilon 1 \ldots \ldots \varepsilon n\}$ *and* $\{x1 \ldots . . , xN\}$ *are independent*
$cov\{\varepsilon i, \varepsilon j\} = 0, i, j = 1, \ldots . . , N\ I \neq j.$
$V\{\varepsilon 1 = \sigma 2, i = 1, \ldots . N$

9. Explain the gradient descent algorithm in detail.

Gradient descent is an iterative form solution for optimizing the cost function, specifically it optimizes minimizing the cost function. It is a solution of order one, implying only first derivative is computed. This is especially useful in cost functions where the second derivative cannot be computed easily. Most linear regression algorithms in python use gradient descent under the hood for minimizing the cost function RSS.
The mathematical representation of gradient descent algorithm is as follows

$$\theta^1 = \theta^0 - \frac{\eta\, \partial}{\partial \theta} J(\theta)$$

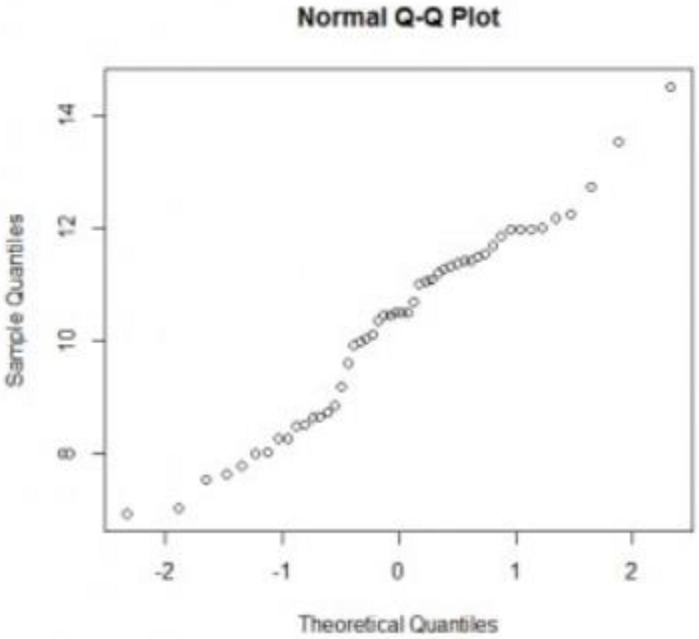
Where η is known as the learning rate and it defines the speed at which we move towards the negative of the gradient. For gradient ascent, the – ve sign is to be replaced with +. Ideally η should be low enough to not miss the minima and high enough to not have unnecessarily long iterations.

The iterative steps for calculating gradient descent is as follows:

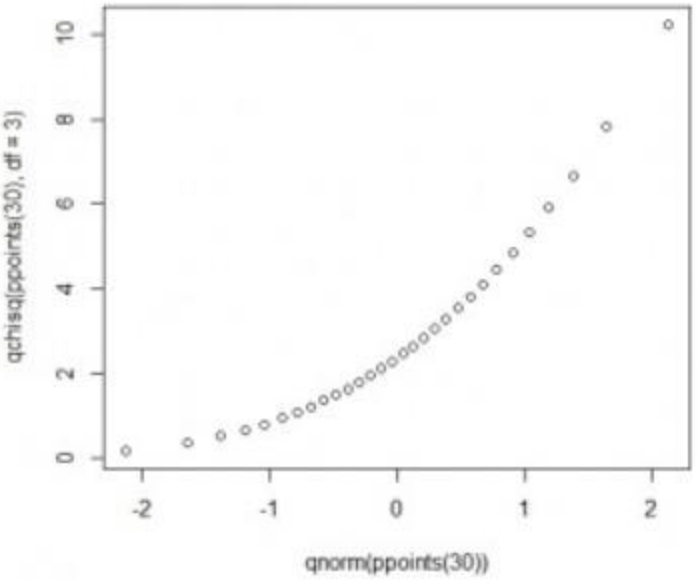
1. Formulate the cost function that we wish to optimize.
2. Obtain the formula for first derivative of the cost function.
3. Take a guesstimate starting value. Choose a learning rate.
4. Find the value of cost function for the above value, and note it as θ^0
5. Find the value of first derivative of cost function for the value and multiply with learning rate, η.
6. Subtract the value obtained in Step 5 from Step 4 as θ^1. θ^1 will be the new value of θ^0.
7. Repeat steps 5-7, until $\theta^1 \sim \theta^0$. Else if maximum number of iterations have been reached, then choose an educated guess starting value and a higher learning rate and repeat steps 4-7.

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a

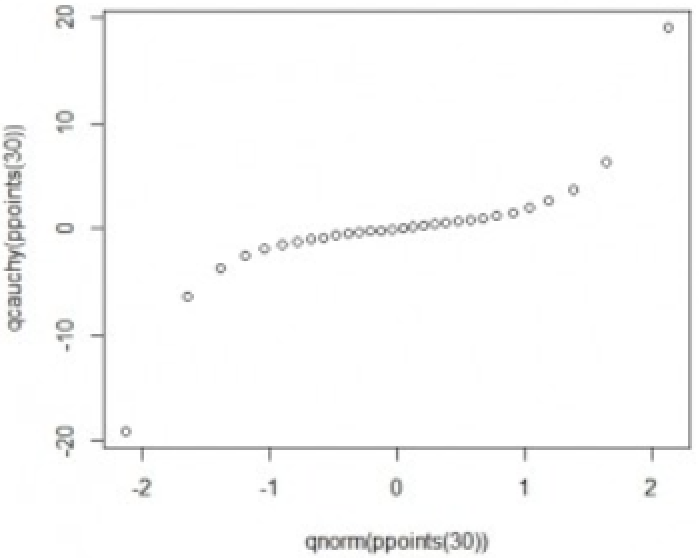
Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can be created for any distribution.



If a plot a chi squared distribution that is skewed right, against a normal distribution, the points would form a curve instead of a straight line.



If we plot a heavy tails versus Normal distribution, points will fall along the line in the middle of the graph but curve off in the extremities.


qcauchy(ppoints(30))
qnorm(ppoints(30))