Question-1:

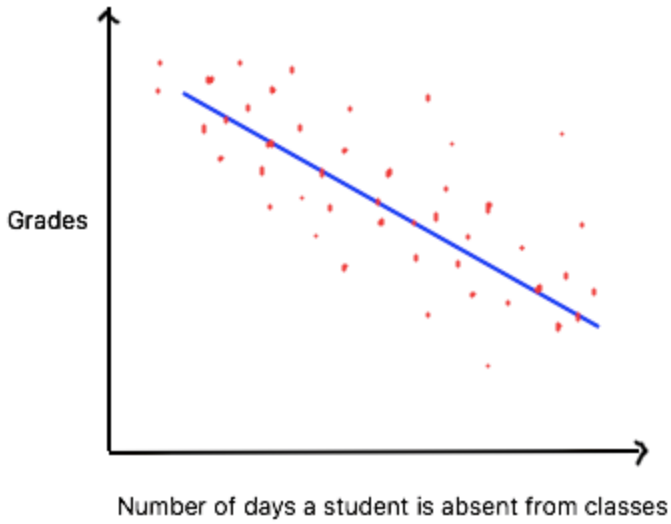
List down at least three main assumptions of linear regression and explain them in your own words. To explain an assumption, take an example or a specific use case to show why the assumption makes sense.

Answer:

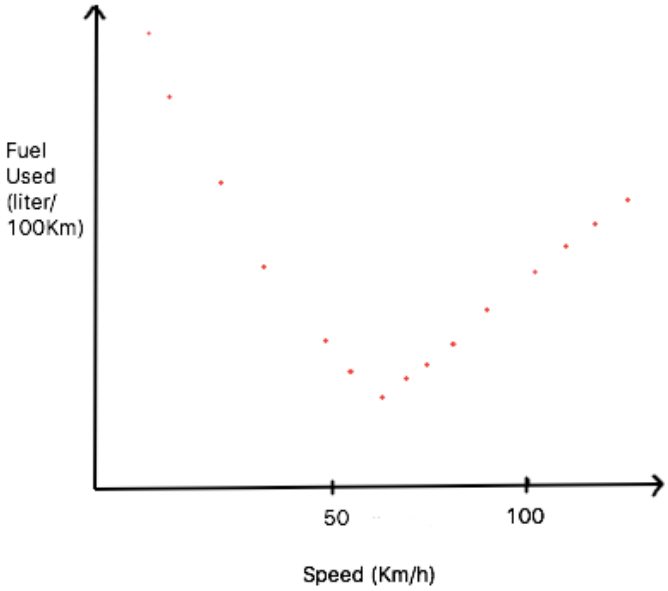
Given below are three important assumptions of linear regression. If the assumptions fail to hold for a given problem, then we cannot apply linear regression to solve that problem.

1. Linear relationship between the dependent and the independent variable.

When a scatterplot between the dependent and the independent variable is plotted, a linear relationship should be observed. The linear relationship can be either a positive or a negative correlation. The plot given below shows how grades drop as students misses more and more classes. The plot clearly shows a linear relationship and the problem of predicting grades based on the number of classes a student has missed can be solved using linear regression.

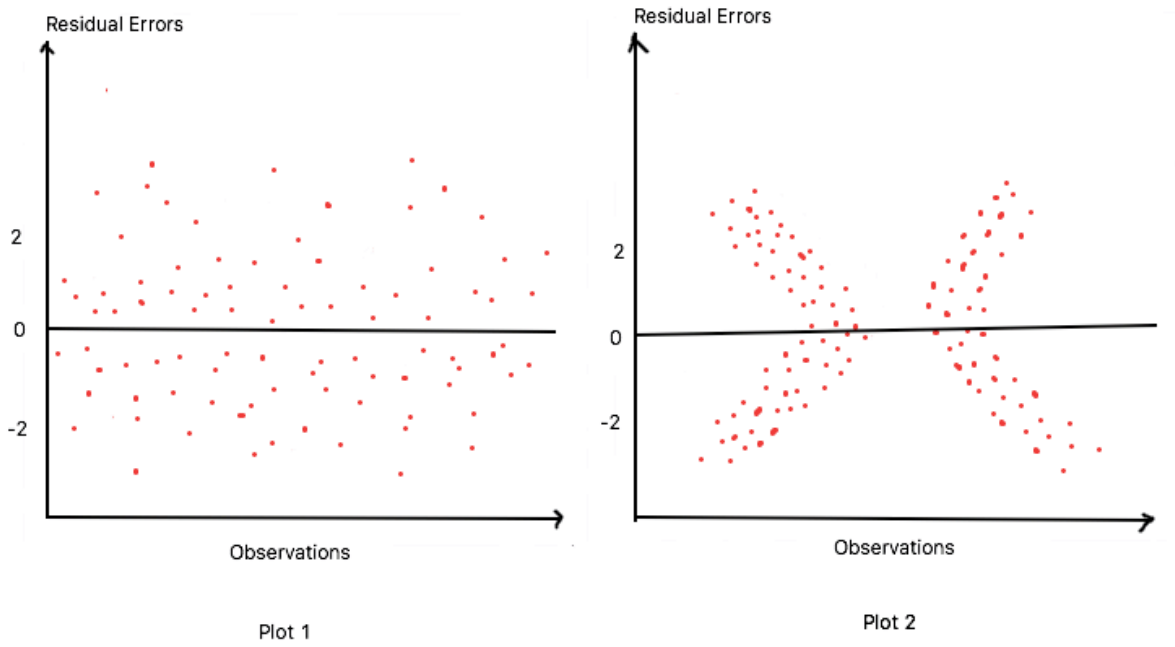


The next plot shows a non-linear relationship. It indicates that as the speed of the car increases from 10 to 60 km/h the amount of fuel consumed by the car decreases. The fuel used by the car starts increasing as the car accelerates beyond 60km/h speed.



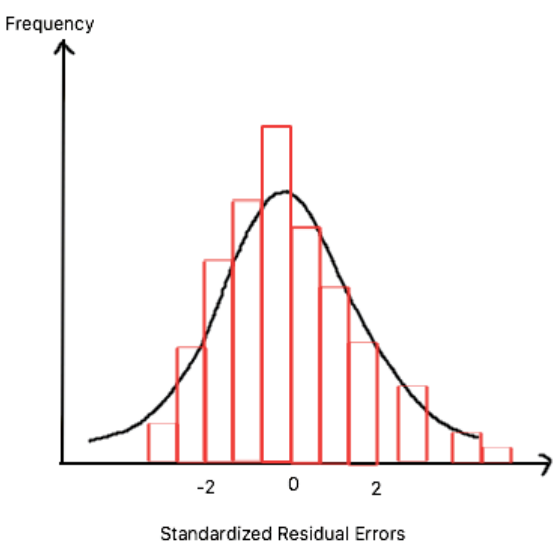
2. Statistical Independence of residual errors.

This assumption applies to the residual errors observed after the linear regression model is trained. The residual error (given by the formula y_observed - y_predicted), when plotted as a time series should indicate a random scattering of data points, with no discernible pattern. In the plots given below, Plot 1 indicates that the linear regression model satisfies this assumption, whereas Plot 2 indicates that there is a certain pattern in the residual errors **which the linear regression model could not explain**. Thus plot 2 violates this assumption and the linear regression model is not correct for plot 2.

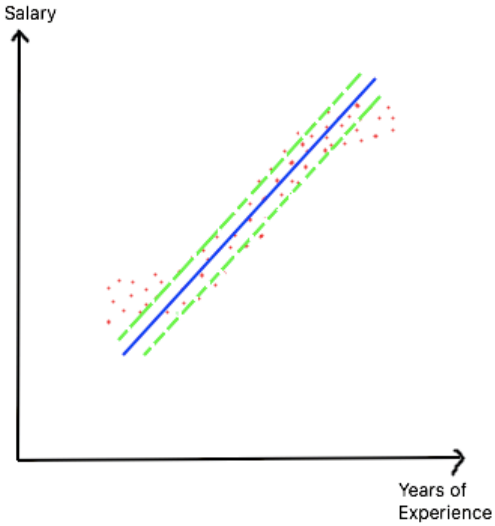


3. Normality of residual error distribution.

The residual errors of a linear regression model should have a normal frequency distribution with peak at 0 as shown in plot 1 below. Plot 2, on the other hand, is non-linear and will not have a normal distribution of residual errors. It's easy to see why. Consider two hypothetical lines (drawn with green) parallel to the linear regression line. These lines indicates one standard deviation of the residual errors on either side of the regression line. For a non-linear curve as in plot 2, we can see that the data points cannot be constrained within such residual boundaries (green lines). For such a curve, the frequency of residual errors will have significant peaks at non-zero values and hence will not be a normal distribution.



Plot 1



Plot 2

Question-2:

Explain at least three regression model evaluation metrics in your own words and compare them.

Answer:

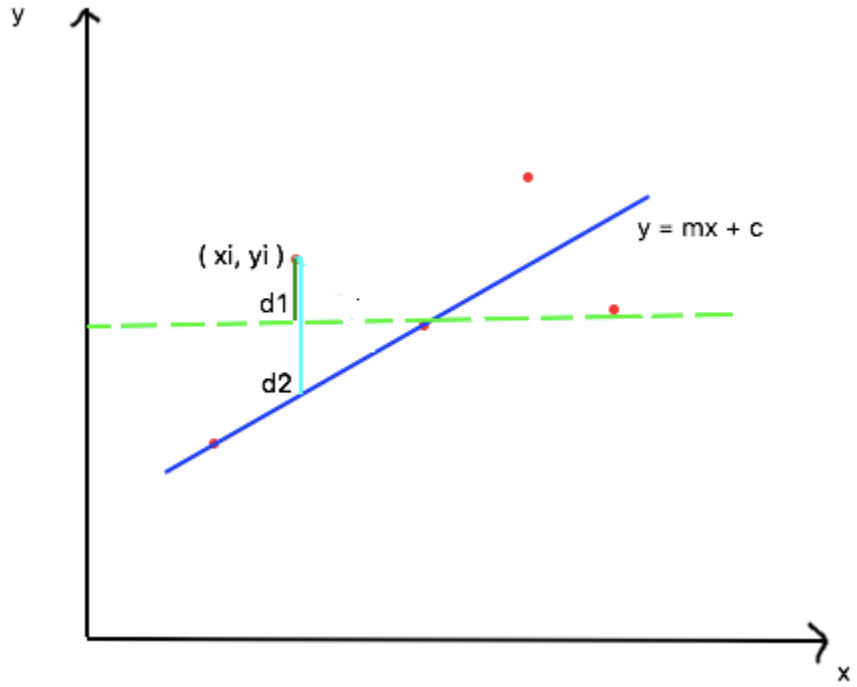
We will discuss the following three regression model evaluation metrics:

1. R-square
2. VIF
3. Root mean square error

For each evaluation metric, we will explain its advantages and disadvantages. At the same time we will compare them with other evaluation metrics whenever required.

R-square

R—square is a statistical measure which indicates how closely a linear regression line fits the data (training set). It is also known as the coefficient of determination and denoted as R^2. A values of R^2 close to 1 indicates that the regression line fits the data very well and a value close to 0 indicates the regression line fits the data poorly. The intuition behind this can be explained clearly using the following diagram.



In the plot above, the red points are the data points and the blue line is a best fit linear regression line. The green dashed line (y_mean) represents the average of the y values. Let's consider a data point given by the coordinates (xi, yi) which is at a distance of d1 from y_mean and d2 from the regression line. Let's call the variance in y as VAR.

$VAR = \Sigma\ (yi - y\_mean)^2$

That is, summing up all the d1-squarres for all the data points.

Now, from the plot, we can see that the error (or squared-error) in measurement for the given coordinate can be written as:

$SE\ (squared\ error) = \Sigma\ (yi - (\ mxi + b\ ))^2$

That is, summing up all the d2-squarres for all the data points.

To find a measure that appropriately defines the quality of fit for the linear regression line, we define a ratio SE/VAR. This ratio indicates the fraction of variance in the data the linear regression model could not explain. We define R^2 as:

$R^2 = 1 - SE/VAR$

For a given set of data points, the variance in the data (VAR) will remain fixed and the squared error (SE) will vary as we rotate the linear regression line gradually to find the best fit during training of the regression model. If the regression line fits the data quite well then SE will be close to 0 and R^2 will be nearly 1. On the other hand, if the regression line does not fit the data well, most of the variance in y will remain unexplained by the regression line, and so SE will be nearly equal to VAR. R^2 in this case will be close to 0. R^2 in this regard can be considered as the amount of variance that is explained by the model. A value close to 1

indicates that the linear regression line can capture most of the variance in the data and hence fits the data points well.

Advantages of R^2

Consider a linear regression model having a high R^2. Furthermore, let's consider that the model is trained on a significant volume of training data that forms an effective representation of all possible data points, which can be generated from a given source. In such a situation, the R^2 would indicate that in the future most data points generated from the same source would be very close to the regression line.

Disadvantages of R^2

R^2 is often a biased estimate which can result in a overfitted model if not used carefully. For example, if a regression model is created based on many predictor variables then the regression line will overfit the training data and the model will have a high R^2. In other words, all the nuances in the training data which are only by chance correlated to the dependent variable (or response variable) will inflate the R^2 and yet result in a poor quality model. It is for this reason, data analysts prefer to use another evaluation metric called adjusted r-square that penalizes the model based on the total number of predictors used to predict the dependent variable. A high R^2 along with a close adjusted R^2 together can determine how good a linear regression model is. In such a situation, another model evaluation metric called Variance Inflation Factor (VIF) can help us determine variables that are unnecessarily boosting the R^2 without significantly contributing in the prediction accuracy of the linear regression model. Using VIF, we can remove such variables and boost adjusted R^2 of the model.

Variance Inflation Factor (VIF)

The variance inflation factor determines the extent of correlation between the predictor variables, or multi-collinearity. When multi-collinearity is present in a model, it is difficult to assess accurately the contribution of predictors to a model. In other words, some variables might have large coefficients although they themselves contribute very less to the accurate prediction by a linear regression model. The reason they have such high coefficient is because their values are automatically boosted due to the presence of other predictors with whom they are correlated.

Let's understand how VIF actually works.

Say, we consider a the equation of a linear regression model as:

$$y = a + bx_1 + cx_2 + dx_3 + ex_4 + fx_5$$

Next, we want to know if, say, the coefficient of predictor x_3 is significant independently or if it is so because x_3 is correlated to other predictor variables. To find this out, we consider this problem as yet another linear regression problem, with the dependent variable as x_3 and all other predictor variables are independent.

$x_3 = p + qx_1 + rx_2 + sx_4 + tx_5$

We then calculate the R^2 value for x_3. A high value of R^2 indicates that x_3 can be accurately predicted from the rest of the predictor variables. Since R^2 is restricted between 0 to 1, we use another metric called the variance inflation factor (VIF) to compare the extent of collinearity present in each predictor variable. Similar to x_3, we will compute VIF for all the variables. VIF of a a variable x_i is given by the formula:

$VIF_i = 1 / (1 - R^2_i)$

When R^2 is quite high, say 0.9, VIF would be 10. For low R^2 like 0.1, VIF would be 1.11. As a rule of thumb, if the VIF of the predictor variables are less than 5, we can assume that the model does not have significant multi-collinearity and is good by this aspect.

Advantages of VIF

(i) VIF provides a tangible idea about how much of the variances of the estimated coefficients of the predictor variables are degraded by multicollinearity. That is, VIF gives us a sense of the severity of multicollinearity present in a linear regression equation. Using VIF, we can improve the adjusted R^2 of a regression model.

(ii) Besides, VIF can capture not only existence of pairwise collinearity between predictor variables but also between sets of predictor variables.

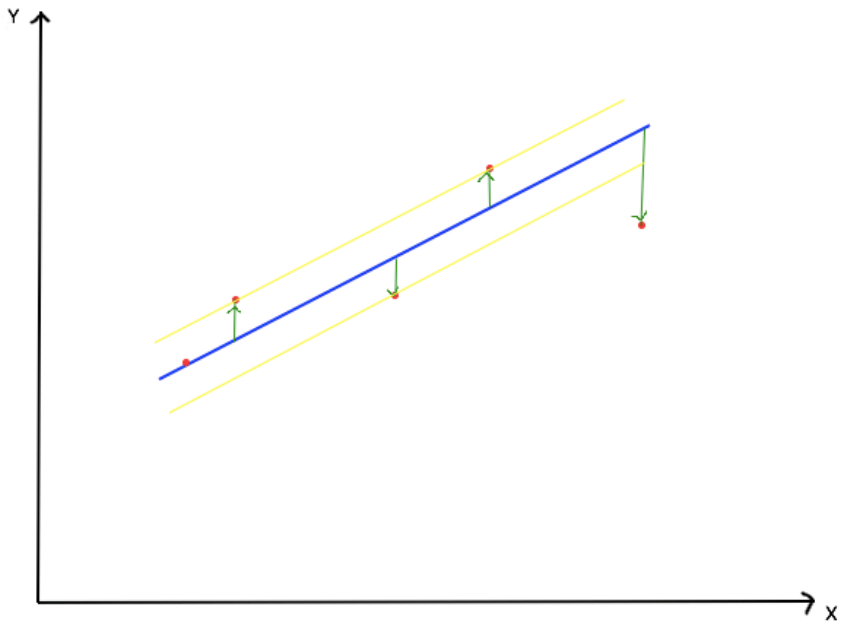
Disadvantages of VIF

(i) VIF does not provide any information about the number of dependencies between the predictor variables.
(ii) The thumb rule VIF >5 or VIF >10 may differ for different linear regression problems.

Root mean square error (RMSE)

The standard deviation of the residuals in linear regression is called the root mean square error. After training a model, we calculate the RMSE using the test data set. This indicates how much of the variance in the test data is unexplained by the linear regression model. A model that has low RMSE indicates that the regression line fits the data points very well.

Let’s understand RMSE using the plot given below:



The red dots are the data points and the blue line is the best fit linear regression line. The green arrows indicate the residual error (RSE) that the model is unable to explain. The residual error of a data point i is given by:

$RSE_i = y_i - y_{predicted}$

The RSE can be both positive or negative. To assess how much residual error a model has, we find the standard deviation of the residual errors. To do this, first we square all the RSE, then find their mean and finally do a square root of the result.

$RMSE = \sqrt{((\Sigma\ y_i - y_{predicted}) / n)}$

To observe how RMSE actually looks in the plot given above, lets observe the two yellow lines drawn parallel to the regression line (blue). These yellow lines are 1 standard deviation away from the regression line. Since we know that the residuals have a normal distribution, most of the data points will be within one standard deviation from the regression line. If RMSE is low the yellow lines will be closer to the blue regression line. In other words, for low RMSE, most of the data points will be very close to the regression line and hence the regression model will be a good fit for the data points.

Advantages of RMSE

R^2 is a relative measure of fit for the linear regression line that varies between 0 to 1. RMSE, however, is an absolute measure of fitness. It indicates the standard deviation of the residuals which is unexplained by the model in the same units as the response variable (y).

Additionally, unlike the Mean Absolute Error (MAE), RMSE takes the square of the residuals while calculating the error. Thus it gives a strong weightage to large errors. This is an important characteristic of RMSE which makes it suitable for solving problems where large errors are especially undesirable.

Disadvantages of RMSE

RMSE is an average of the errors observed in the test data set. This could be an issue if there are a large number of errors in the test data set, which are individually not large in magnitude. RMSE will average out all the errors in the test dataset and hence will not indicate any issue with the model. This could be problematic for certain business situations.

RMSE treats all the errors in the same way and penalizes all the errors for all values of the response variable (y) equally. This may not be an accurate reflection of the real world. For example, consider a business problem where we are okay to accept large number of errors of small magnitude for lower values of y (near to 0), but we do not expect errors of large magnitude for these small y values. Also, for large y values, we are okay to accept errors of large magnitude but we expect them to be few in numbers. Such a behaviour cannot be captured using RMSE.

R^2 is a more interpretable measure of fitness with values close to 1 indicating a good fitness for the regression line. RMSE, although, not so interpretable, provides more information than R^2 by explaining how much the data points are scattered away from the regression line (or the standard deviation of the residuals).